

AD-A060 786

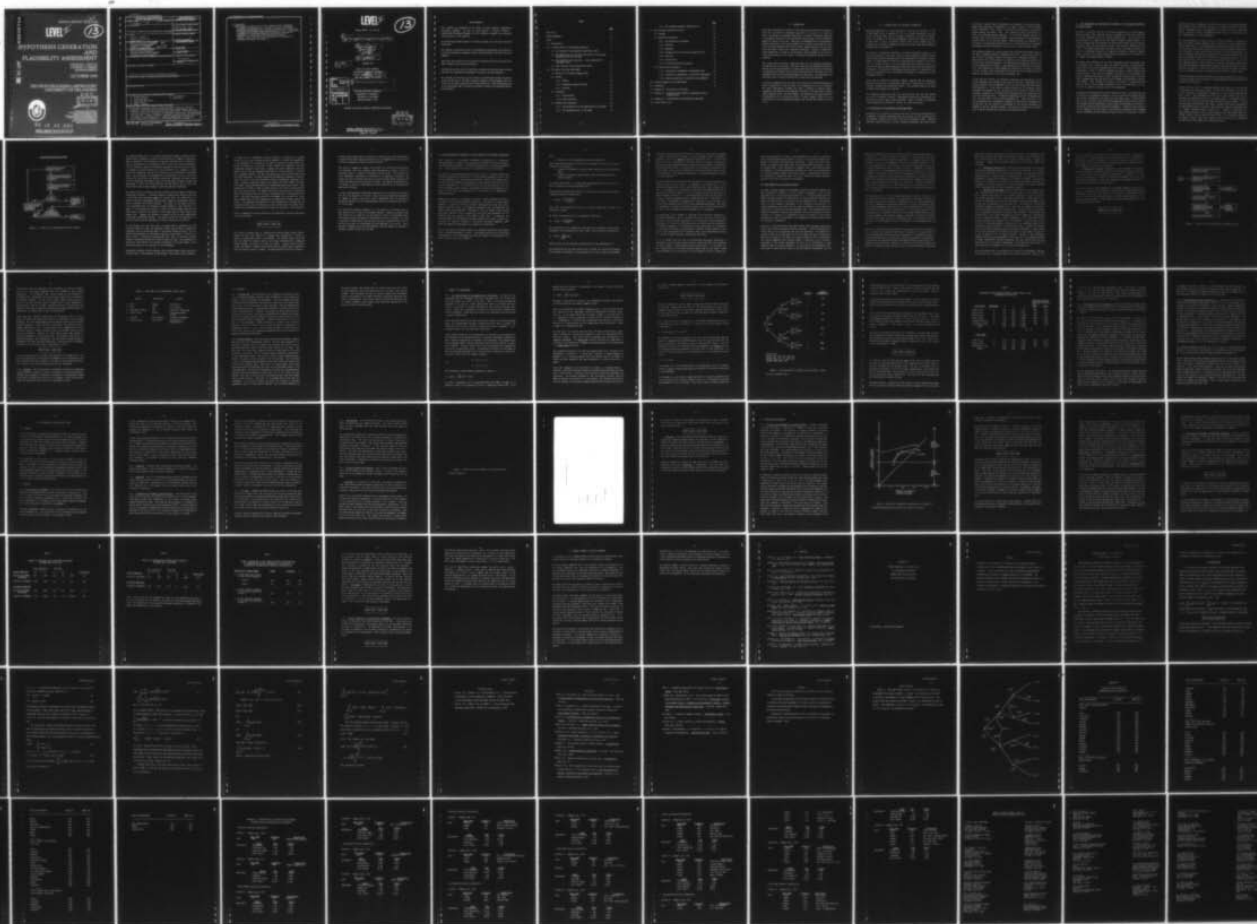
OKLAHOMA UNIV NORMAN DECISION PROCESSES LAB  
HYPOTHESIS GENERATION AND PLAUSIBILITY ASSESSMENT.(U)  
OCT 78 C F GETTYS, S D FISHER, T MEHLE

F/G 5/10

UNCLASSIFIED

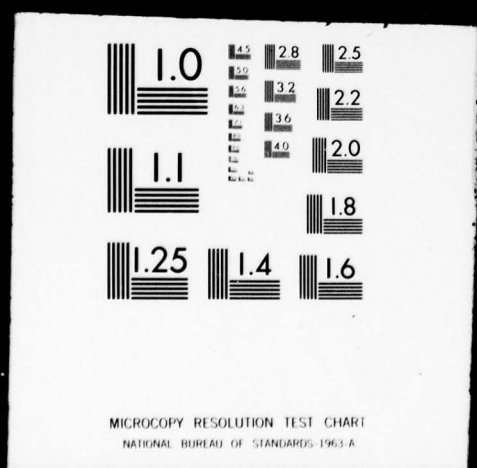
N00014-77-C-0615  
NL

1 OF 2  
AD  
A080788



1 OF 2

AD  
A060786



AD A060786

ANNUAL REPORT TR <sup>date</sup> 15-10-78

LEVEL <sup>II</sup>

13

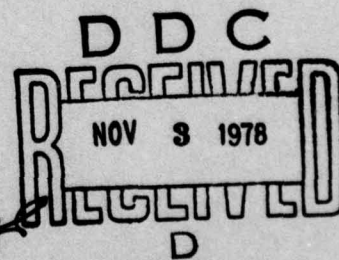
# HYPOTHESIS GENERATION AND PLAUSIBILITY ASSESSMENT

CHARLES F. GETTYS  
STANLEY D. FISHER  
THOMAS MEHLE

OCTOBER 1978

DDC FILE COPY

DECISION PROCESSES LABORATORY  
UNIVERSITY OF OKLAHOMA



OFFICE OF NAVAL RESEARCH  
CONTRACT NUMBER N00014-77-C-0615  
WORK UNIT NUMBER NR197-040  
REPRODUCTION IN WHOLE OR IN PART IS  
PERMITTED FOR ANY PURPOSE OF THE  
UNITED STATES GOVERNMENT  
APPROVED FOR PUBLIC RELEASE;  
DISTRIBUTION UNLIMITED.

78 10 26 001

ORIGINAL CONTAINS COLOR PLATES: ALL DDC  
REPRODUCTIONS WILL BE IN BLACK AND WHITE.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER TR 15-16-78 <i>late</i>	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  HYPOTHESIS GENERATION AND PLAUSIBILITY ASSESSMENT		5. TYPE OF REPORT & PERIOD COVERED Annual-15 Aug., 1977- 15 Aug. 1978
7. AUTHOR(s) Charles F. Gettys Stanley D. Fisher and Thomas Mehle		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decision Processes Laboratory <i>New</i> <del>Department of Psychology</del> University of Oklahoma Norman, Oklahoma 73019 <i>H10918</i>		8. CONTRACT OR GRANT NUMBER(s)  N00014-77-C-0615 <i>aw</i>
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Engineering Psychology Programs Code 455, 800 N. Quincy Street Arlington, Virginia 22217		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS  NR 197-040
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE 15 Oct 1978
		13. NUMBER OF PAGES
		15. SECURITY CLASS. (of this report)  Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release, distribution unlimited		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) 1) decision theory 2) human inference 3) hypothesis generation 4) memory search 5) Bayes' theorem 6) plausibility 7) heuristics		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A hypothesis generation model is described which consists of two sub-processes. Hypotheses are retrieved from memory using several data as retrieval cues in the hypothesis retrieval sub-process. These hypotheses are then evaluated by a plausibility assessment sub-process. Two experiments are described. A memory retrieval experiment examined hypothesis retrieval from memory using multiple data. A memory-tagging model is described which predicts the probability of multi-data hypothesis retrieval.		

DD FORM 1 JAN 73 1473


EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-014-6601

Unclassified

78 10 26 001  
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Continued

Performance in this task was poor; subjects rarely generated an adequate hypothesis set. A second plausibility assessment experiment was performed where subjects estimated the plausibility of specified hypotheses using varying amounts of data. Plausibility assessments for specified hypotheses were usually extreme in comparison to the posterior odds calculated by Bayes' theorem. This result was also attributed to deficiencies in hypothesis retrieval from memory.



Unclassified

# LEVEL II

13

ANNUAL REPORT TR 15-10-78

6 HYPOTHESIS GENERATION AND PLAUSIBILITY ASSESSMENT.

10 by  
CHARLES F. GETTYS,  
STANLEY D. FISHER  
THOMAS MEHLE

11 15 OCTOBER 1978

12 97P.

PREPARED FOR

OFFICE OF NAVAL RESEARCH  
ENGINEERING PSYCHOLOGY PROGRAMS  
CONTRACT NUMBER 15/111 N00014-77-C-0615  
WORK UNIT NUMBER NR 197-040

ADDITIONAL TO	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY GROUP	
DEL.	AVAIL. and/or SPECIAL
A	

9 Annual rept. 15 Aug 77-15 Aug 78,

DECISION PROCESSES LABORATORY  
DEPARTMENT OF PSYCHOLOGY  
UNIVERSITY OF OKLAHOMA  
NORMAN, OKLAHOMA 73019

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

DDC  
RECEIVED  
NOV 3 1978  
RECEIVED

ORIGINAL CONTAINS COLOR PLATES: ALL DDC  
REPRODUCTIONS WILL BE IN BLACK AND WHITE.

420 928

### Acknowledgments

This research was supported by the Office of Naval Research, Engineering Psychology Programs. We wish to thank Martin A. Tolcott, Director, Engineering Psychology Programs, for his insightful comments and conceptual contributions to this research.

Carol Manning performed several of the data analyses, and critically read this manuscript.

Dale Umbach, and Bradford Crane of the Mathematics Department participated in the proofs presented as appendix A and W. Alan Nicewander critically read this appendix.

Chuck Rice and Jim White of the University Computing Services provided invaluable software and consulting services.

Tom Harrison and the staff of the Office of Grants and Contracts have provided editorial and secretarial assistance for which we are most grateful.

Our interest in this topic can be traced to conversations between the senior author and Mel Moy of the Navy Personnel Research and Development Center on the topics of threat detection and crisis prevention. These discussions lead to the conclusion that it would be profitable to model the hypothesis generation process.

## INDEX

	<u>Page</u>
Form 1746	i
Acknowledgements	iii
Index	iv
1.0 Introduction	1
2.0 A process model for hypothesis generation	2
2.1 An overview of the hypothesis generation model	2
2.2 How hypotheses are retrieved from memory- the proposed hypothesis retrieval process	4
2.3 How hypotheses are evaluated- a dual plausibility assessment model	10
2.4 Past research on the plausibility model	13
3.0 The memory retrieval experiment	17
3.1 Details of the memory tagging model	17
3.2 Method	18
3.2.1 Design	18
3.2.2 Hypothesis generation tasks	18
3.2.3 Subjects	19
3.3 Procedure	20
3.3.1 Instructions	20
3.3.2 Data collection	20
3.4 Results and discussion	22
3.4.1 The calculation of the predictions of the model	22
3.4.2 The goodness-of-fit of the model	26

	<u>Page</u>
3.4.3 The minimally-adequate hypothesis set	27
4.0 The veridical plausibility study	29
4.1 Purpose	29
4.2 Method and procedure	29
4.2.1 An overview of the method	29
4.2.2 Subjects	30
4.2.3 Apparatus	30
4.2.4 Estimation of the veridical plausibilities	30
4.2.5 Problems	31
4.2.6 Instructions	32
4.2.7 The data collection procedure	32
4.3 Results and discussion	34
4.3.1 Plausibility assessment of hypothesis sets	34
4.3.2 Plausibility assessment of individual hypotheses	37
4.3.3 Ordinal properties of plausibility assessments	39
5.0 General summary of both experiments	41
6.0 References	43
7.0 Appendix A: Derivation of the model	
8.0 Appendix B: Predicted versus emperical hypothesis recall probabilities	
9.0 Appendix C: Problems used in plausibility experiment	
10.0 Distribution list	

## 1.0 INTRODUCTION

This report describes the results of the first year of an effort to develop a model for the process of hypothesis generation. Our goal is to study and model the process of hypothesis generation to provide information pertinent to the general process of decision-problem structuring, in which hypothesis generation plays a vital role. Hopefully, an understanding of the hypothesis generation process of the decision maker will be useful in two general areas. First, decision analysts can profit from an understanding of the heuristic rules that their clients use to generate hypotheses. Second, if human hypothesis generation is found to be deficient, then knowledge of the cause of these deficiencies will be useful if hypothesis generation aiding is to be provided.

This report first describes a tentative model for the hypothesis generation process which we are evolving. This model separates the hypothesis generation process into two sub-processes; one sub-process which describes the retrieval of hypotheses from memory, and a second sub-process which describes how those hypotheses that are retrieved from memory are evaluated. Two experiments are described each of which is devoted one of these two sub-processes.

It should be emphasized at the onset that because this hypothesis generation model is ambitious in scope that many of its assumptions are as yet untested. Our basic experimental strategy has been to attempt to identify those assumptions and questions that seem most critical to the model and to address these questions first. For this reason, this version of our model should not be considered to be our definitive effort. We have, however, found it to be a useful guide to our thinking and research, and we hope that others will find it so.

## 2.0 A PROCESS MODEL FOR HYPOTHESIS GENERATION

Hypothesis generation is a vital precursor to the decision process. If the Decision Maker fails to generate all of the relevant hypotheses, any subsequent decisions made with an incomplete hypothesis set may be inappropriate. For example, if a physician diagnoses a patient as having one of four diseases when, in fact, the patient suffers from a fifth disease, then the coherency of the medical diagnosis process has broken down, and subsequent treatment may be ineffectual.

The process of hypothesis generation is poorly understood because of the paucity of research on this topic. Much of the psychological research investigating human decision making parallels the development of normative decision models. As these models are basically algorithms which operate on the structure of a decision model, it is not surprising that questions having to do with how the Decision Maker generates these structures have been postponed. Furthermore, the problem of developing normative models for hypothesis generation has apparently been intractable and probably will remain so for the foreseeable future.

Recent work in cognitive psychology, however, suggests that the hypothesis generation process can be profitably modeled by a combination of decision-theoretic concepts and descriptive theory. The hypothesis generation model discussed here employs this approach.

This hypothesis generation model has been evolving over the last several years, and is tentative in nature. Experimental results are incorporated into the model as soon as they become available. Hence, the model is constantly being elaborated and modified to account for these new results.

### 2.1 An Overview of the Hypothesis Generation Model

This model is to be applied in those situations in which the decision maker is attempting to generate hypotheses that will account for the available data. For example, a physician has data from various diagnostic tests. When the physician inspects the data, various diseases (hypotheses) may come to mind.

The process of generating new hypotheses can be modeled in the following way. Assume that hypotheses are generated by a highly specific recursive memory search (Shriffrin, 1970), which is controlled and guided by an executive process (Newell and Simon, 1972). This executive process initiates, directs, and terminates memory searches. It is further assumed that one important input to the executive process is the plausibility of any hypothesis currently held by the Decision Maker. Memory searches are assumed to be initiated if no hypotheses currently exist, or if the plausibility of hypotheses already retrieved from memory is so low as to require further search.

The hypothesis generation process begins when the executive becomes aware of the need to generate possible explanations for data. The executive directs and controls the memory search and plausibility assessment processes. A memory search for new hypotheses is initiated by the executive based on the data that are currently available. The memory search process may retrieve one or more hypotheses from memory. These hypotheses are returned to the executive. Each hypothesis is then assessed for plausibility in light of the data on hand. Finally, those hypotheses that survive the test of plausibility are added by the executive to the current hypothesis set.

If new data are received, the executive reassesses the plausibility of the current hypothesis set taking the new data into account. Hypotheses may be dropped from the current hypothesis set at this time because new data renders them implausible. If the total or global plausibility of the entire current hypothesis set becomes too low, the executive will initiate a further search of memory attempting to find additional hypotheses that are consistent with both the old and new data. If these new hypotheses survive the plausibility test, they are added to the current hypothesis set. Thus, the size of the current hypothesis set increases or decreases as new data are incorporated in the process, and the identity of hypotheses in the current hypothesis set changes as new data becomes available. The process is recursive in the sense that it may be repeated each time new data become available.

## 2.2 How Hypotheses are Retrieved from Memory - The Proposed Hypothesis Retrieval Process

Generally, we consider the act of retrieving hypotheses to involve semantic memory (Tulving, 1972) since this process is thought to involve the retrieval of factual information from the long-term memory store. At the present we have decided not to adhere to any one specific semantic memory theory. Instead, concepts have been adopted from both network models such as Anderson and Bower (1973) and set-theoretic models such as Smith, Shoben and Rips (1974). The reason behind such a decision is that our primary interest does not lie in providing evidence for or against any specific theory of memory. Rather, our goal is to study the generation of hypotheses in a way which can be adapted to different models of memory.

We suppose that hypotheses are retrieved from memory using the relevant data as retrieval cues. Based on data, the subject conducts a highly-specific memory search to retrieve hypotheses which can account for the available data. Consequently, direct or indirect linkages must exist in memory between data and hypotheses. There may be accessible information in the linkages themselves. We assume that some of these linkages are associational in nature; these linkages exist when the data directly suggest the hypothesis in associative memory. Other linkages are indirect, or mediated. In these cases the data are used to retrieve some intermediate event or variable which, in turn, serves as an implicit retrieval cue for the hypotheses.

We suppose that hypotheses are rarely invented "de novo", but rather that the availability of a hypothesis depends on its prior existence in memory and upon the content, organization, and structure of the memory store. Decision Makers must be able to exploit their factual knowledge which specifies the relationship between data and hypotheses; usually data and hypotheses are directly or indirectly related by facts retrieved from memory. In addition they should be able to engage in accurate inductive and deductive reasoning as the linkages between hypotheses and data often are chains of indirect reasoning. Finally,

they must be able to recognize the similarities and the differences between their present situation and past situations. We assume, therefore, that their factual information store, the organization of this store, their reasoning ability, and their ability to generalize from the past to the present are major determinants of their performance.

The actual hypotheses retrieval process of Decision Makers is complicated when the Decision Maker possesses data occurring in a novel combination. Ideally their goal is to retrieve hypotheses which are consistent with all of the known data; in practice they may settle for less than perfect consistency. First consider the case where N data are known to the Decision Maker. An actual hypothesis generation task employed in the present research will be used as an example to make the discussion as concrete as possible. In this task subjects were given the notable products and industries of one of the fifty American States. Their task was to retrieve from memory States which they believed might have generated the products and industries given as data. They were instructed to search their memory for States which were consistent with all the data given and to respond with any State that came to mind.

Considering States retrieved from memory as possible hypotheses, our goal is to develop a model for the retrieval of these hypotheses from the products and industries data provided.

Suppose that the task is to generate hypothesized States for the following three products and industries 1) Beef, 2) Fish, and 3) Aerospace. The concept of each datum may be represented in memory as a node or point (Anderson and Bower, 1973). The hypotheses that are yet to be retrieved can also be similarly represented as points in this memory space. If subjects had only one datum as a retrieval cue then their task would be much easier. Suppose they were given the retrieval cue of Beef. Subjects who actually used beef as a single-datum retrieval cue gave a wide variety of States as responses such as: Kansas, Oklahoma, Texas, Colorado, etc. Similarly, using fish as a single-datum retrieval cue led to many seaboard, and some inland States including: Maine, Florida, Texas and many others. Since aerospace industries are located in many states their responses were also varied, but frequently included the two main NASA sites, Texas and Florida.

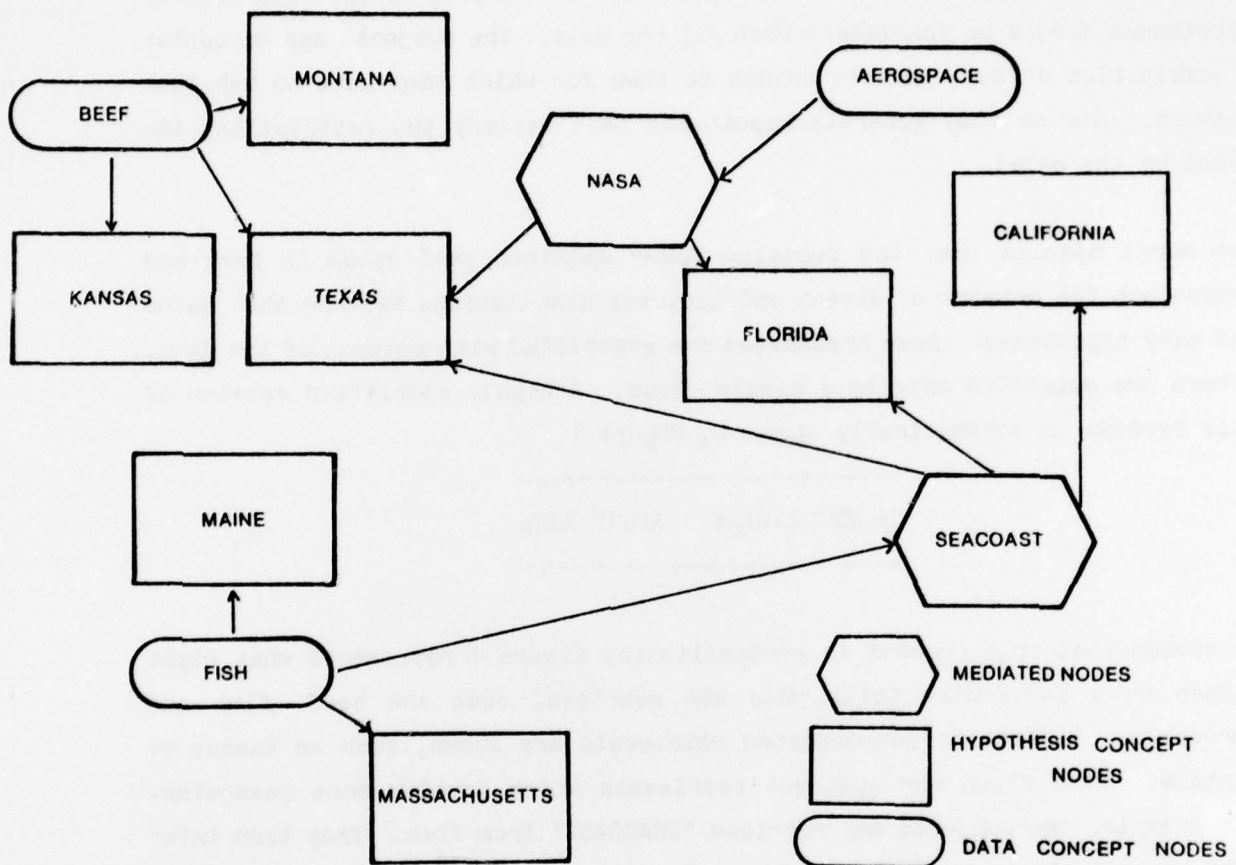


Figure 1. A highly - simplified representation of the associational network between hypotheses and data. Both direct and mediated associations are shown.

Multiple-data retrieval is more difficult for the subject because the desired hypotheses should be consistent with all the data. The subjects may encounter a combination of data that is unique to them for which they have no habitual answers. How do they generate hypotheses that satisfy the restrictions imposed by the data?

Our model assumes that the Decision Maker examines each datum in turn and traces out the network of direct and indirect associations between that datum and many hypotheses. Some hypotheses are associated with several of the data, others are suggested only by a single datum. A highly simplified version of this process is schematically shown in figure 1.

-----  
INSERT FIGURE 1 ABOUT HERE  
-----

We assume that this process is probabilistic; figure 1 represents what might happen on a particular trial when the retrieval cues are beef, fish and aerospace. Direct, or non-mediated retrievals are shown, such as Kansas or Montana. Also shown are indirect retrievals which involve more reasoning. For example, the subjects may retrieve "SEACOAST" from fish. They then infer that coastal states probably are known for the production of fish, leading to the retrieval of Texas, Florida, and California from the category of coastal states.

Direct retrievals are assumed to be habitual responses to data which have become automatic through repetition. For example, Texas might be considered a habitual response to beef. In terms of Schiffman and Schneider (1977), such direct retrievals would constitute an example of "automatic processing". In such a case, the datum should automatically activate a hypothesis node without any attention being allocated for such a retrieval. Direct retrievals should thus occur as a function of the amount of practice or repetition which has been devoted to the storage of the direct association between a given hypothesis and any relevant data. In addition, direct retrievals should require a minimum of conscious processing activity and attentional demands. However,

## HYPOTHESIS RETRIEVAL MODEL

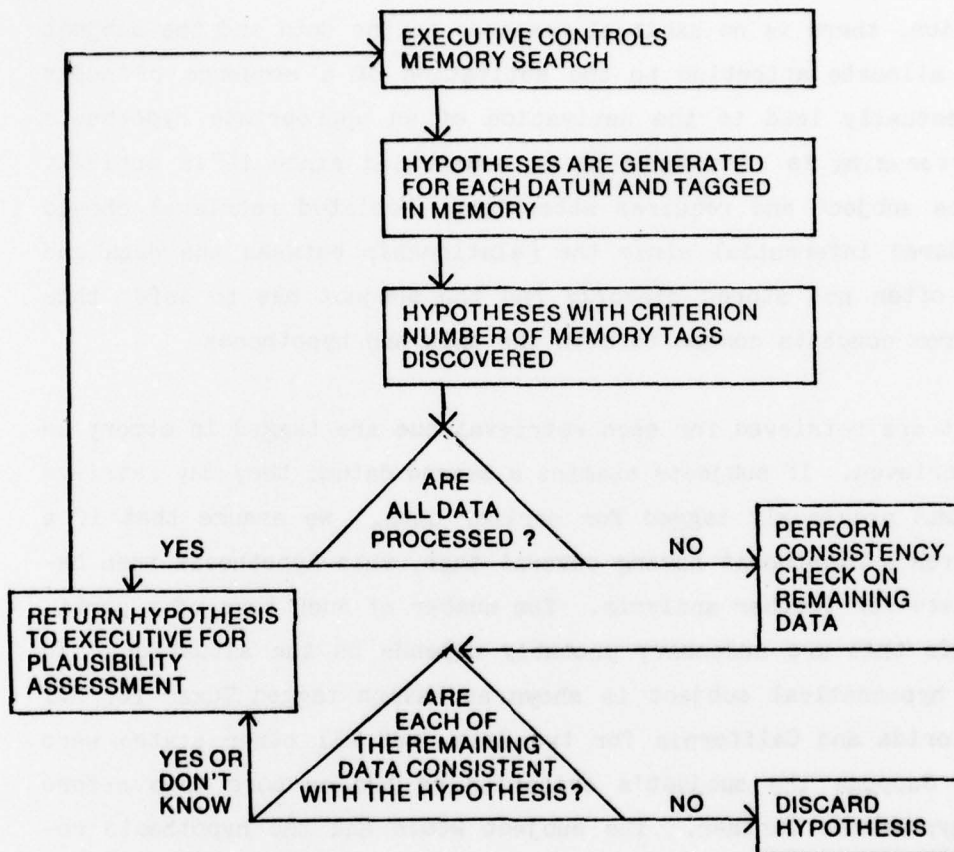


Figure 2. A model of the hypothesis retrieval process.

in situations where there is no direct association between a datum and hypotheses stored in memory, an indirect retrieval may also suggest a hypothesis. What we consider to be an indirect or mediated retrieval can be considered a case of what Schifffrin and Schneider (1977) called "controlled processing". In this situation, there is no habitual response to the data and the subject is assumed to allocate attention to the activation of a sequence of nodes which will eventually lead to the activation of an appropriate hypothesis node. Such processing is considered to be controlled since it is actively directed by the subject and requires attention. Mediated retrieval should also be considered inferential since the relationship between the data and hypotheses is often not stored directly and the subject has to infer this relationship from concepts common to both the data and hypotheses.

The states that are retrieved for each retrieval cue are tagged in memory as having been retrieved. If subjects examine a second datum, they may retrieve a State that was previously tagged for earlier data. We assume that if a subject retrieves a hypothesis having several tags, this hypothesis then becomes a candidate for further analysis. The number of such tags of a particular hypothesis that are necessary probably depends on the situation. In figure 1, the hypothetical subject is shown as having tagged Texas for all three data; Florida and California for two data, and all other states were tagged once. Suppose the subject's criterion is two or more tags before processing a hypothesis further. The subject would end the hypothesis retrieval process with the hypotheses of Texas, Florida and California.

If the subject still has more data to process when a hypothesis has the criterion number of tags, the hypothesis search process may be temporarily suspended, and the subject may begin a consistency checking process. We will discuss the mechanism of consistency checking in more detail in Section 2.3 which follows. Consistency checking may occur when the retrieval process locates a hypothesis having the criterion number of tags. This hypothesis is returned to the executive for further processing, the Decision Maker is now consciously aware of the candidacy of the hypothesis.

Consistency checking involves making a more limited memory search using two retrieval cues - the datum and the hypothesis. This search is more specific

in nature than in hypothesis retrieval because it consists of a memory search for relational information which supports or refutes the hypothesis. As the hypothesis is now known and is used in conjunction with the datum as a second retrieval cue, the search is more limited and attempts to retrieve information relevant to both cues. Suppose that three additional data were added to beef, fish and aerospace which were 4) citrus products, 5) tourists and 6) cypress products, and that the subject has just retrieved Texas from the first three data. He could now search his memory to determine if Texas is noted for citrus, and would probably retrieve information consistent with Texas. Another consistency check would compare the fourth datum, tourists, with Texas. Upon checking tourists an interesting complication arises. Most subjects do not think of Texas as being noted for tourists but Texas certainly must have some tourists. We imagine that Texas would survive a consistency check if the subject does not retrieve information that is inconsistent with Texas. The datum "cypress products" is an example of a retrieval cue where many subjects did not have information in memory which linked cypress and Texas. It is tentatively assumed that if such information is lacking that the hypothesis usually will be retained, but that a datum which is clearly inconsistent with a hypothesis usually will cause the hypothesis to be discarded.

Figure 2 summarized the major features of the hypothesis retrieval subprocess of the total model.

-----  
INSERT FIGURE 2 ABOUT HERE  
-----

The concept of memory tags is consistent with modern memory theory (Smith, Shoben and Rips, 1974) where it is assumed that information is stored in memory in associated clusters of attributes. For example, facts associated with the "Bear D" aircraft, such as its flight characteristics, range, and sensor systems, are stored in a "Bear D" cluster. A memory tag is an additional fact which is added to this cluster and is a marker that a hypothesis node was recently active in memory. As some hypothesis nodes will acquire

multiple tags because they are retrieved for several data, those tags provide a mechanism for identifying hypotheses which are suggested by the collection of data, an essential feature for any such model.

This notion of tagging by recency of activation is consistent with Schifffrin and Schneider (1977) who consider short-term memory as being a series of activated nodes in a network of otherwise inactive nodes. If a concept node which represents a hypothesis is retrieved it is temporarily activated. We assume that this activation can serve as a tag or marker which indicates that a given hypothesis has been related to a datum. This memory tag assumption is an attempt to provide a psychological process explanation for multiple-data hypothesis retrieval.

A single datum generally evokes many hypotheses; most of which are inconsistent with the remainder of the data. While we are assuming that hypotheses are tagged in memory, many other similar mechanisms could be proposed to accomplish the goal of retrieving hypotheses which are consistent with most or all of the data.

The consistency checking assumption is a new addition to the model, and for this reason is not a topic in the present research. However, we believe that this feature of the model will be particularly useful for describing situations where the Decision Maker has large amounts of data to process. In these situations it seems logical that he would retrieve a hypothesis from part of the data and check its consistency with the remainder of the data. Such a heuristic strategy would save considerable time and thought as compared to exhaustive memory searches on all data.

### 2.3 How Hypotheses are Evaluated - A Dual Plausibility Assessment Subprocess

When a hypothesis is retrieved from memory its plausibility is evaluated. Some hypotheses are found to be plausible and are added to the set of hypotheses that the Decision Maker is currently entertaining (the current hypothesis set). Other hypotheses are found to be implausible and are discarded.

The plausibility assessment process is dual. First, it is used to assess each individual hypothesis to decide if it is sufficiently plausible to use in the current hypothesis set, as previously mentioned. Second, the plausibility of the individual hypotheses are cumulated to yield a plausibility for the entire current hypothesis set. The plausibility of the current hypothesis set controls the memory search process. Hypothesis retrievals from memory are assumed to cease when the plausibility of the entire set of hypotheses is sufficiently high.

Modeling the plausibility estimation process, unlike hypothesis retrieval, can profit from decision-theoretic constructs. Our goal is to apply Bayes' Theorem to a process that has different characteristics than the typical Bayesian inference task. In a Bayesian inference task, the hypotheses are usually known and enumerated. Furthermore, Decision Makers usually assume that their hypothesis set is exhaustive; that they have enumerated all hypotheses that are possible in the light of their data. In practice, the Decision Makers may deliberately neglect possible but highly unlikely hypotheses, or they may introduce a "catch-all" hypothesis (Edwards, 1966) which then makes the hypothesis set exhaustive by definition.

On the other hand, the Decision Maker in a hypothesis generation task begins with no specific hypothesis in mind. The process starts with a "yet-to-be-enumerated" set of all hypotheses,  $H$ , which are assumed to be possible for the data,  $D$ , for all  $H_i$  in  $H$ ,  $P(D|H_i) > 0$ .

Let:

$H = \{\text{finite set of all hypotheses for which } P(D|H_i) > 0.\}$

When hypotheses are generated by the Decision Maker the set  $H$  can be partitioned into two subsets:

$C = \{\text{subset of hypotheses currently under consideration by the Decision Maker.}\}$

$\bar{C} = \{\text{subset of hypotheses not currently under consideration by the Decision Maker.}\}$

It follows that  $C \cup \bar{C} = H$ , as any given hypothesis must be in  $C$  or in  $\bar{C}$ .  $C$  will be termed the "current hypothesis set".

The posterior plausibility,  $P(H_i|D)$ , is defined over the entire set  $H$ , and is a Bayesian posterior probability:

$$1) \quad P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)},$$

where  $P(D)$  is calculated for all  $H_i$  in  $H$ , and  $D$  stands for a datum, or a collection of data.

We define the plausibility of  $C$  using Bayes' theorem as:

$$2) \quad P(C|D) = \frac{P(C) P(D|C)}{P(D)}.$$

The term  $P(C|D)$  is the probability that one of the hypotheses in the current hypothesis set has generated the data, and it can be readily calculated as:

$$3) \quad P(C|D) = \sum_{H_i \in C} P(H_i|D),$$

where  $P(H_i|D)$  are the posterior plausibilities of the hypotheses in  $C$ .

We deliberately use the term plausibility to remind the reader that although the concepts are Bayesian, our application is to the process of hypothesis

generation, which requires the evaluation of quantities not usually evaluated in Bayesian probabilistic inference. The Decision Maker's goal in hypothesis generation task is to generate plausible hypotheses for C. The Decision Maker's goal in a probabilistic inference task is to evaluate the posterior probabilities of hypotheses which have been previously generated. In hypothesis generation tasks,  $P(C|D)$  is initially zero, as the subset of hypotheses currently under consideration is empty until the Decision Maker starts hypothesis generation. After successful completion of a hypothesis generation task,  $P(C|D)$  may approach 1.0.

The Decision Maker has two alternatives available to deal with a probabilistic inference task for which  $P(C|D) < 1$ . One approach is to assume  $P(C|D) = 1$ , thus assuming that hypotheses not in C may be neglected with a small probability of error. For example, in a coin-flipping task, the Decision Maker may choose to neglect the unlikely alternative of a coin landing on edge, and may choose to define  $P(\text{Heads}) + P(\text{Tails}) = 1$ . Another approach is for the Decision Maker to include a catch-all hypotheses in C, in which case  $P(C|D) = 1$  and  $P(\bar{C}|D) = 0$  by definition. The emphasis in probabilistic inference is on the relative size of the  $P(H|D)$  probabilities for the various enumerated hypotheses. The catch-all hypothesis is usually of secondary interest, and is often included solely so that it can be claimed that the hypothesis set is exhaustive.

All hypotheses can be thought to originate from "catch-all" hypothesis in hypothesis generation, and the primary focus is on the size of the probabilities of the enumerated hypotheses in C in comparison to the un-enumerated hypotheses in  $\bar{C}$  which can be conceptualized as a giant catch-all hypothesis. Our purpose, therefore, in introducing the term plausibility is to emphasize the possible differences in psychological processes between hypothesis generation tasks and probabilistic inference tasks, while simultaneously noting the utility of Bayes' theorem in describing both tasks.

If it is assumed that the cost of entertaining additional hypotheses is negligible, then the "Ideal Decision Maker" should entertain all hypotheses for which  $P(D|H) > 0$ . This logically implies that  $P(C|D) = 1.0$ . Therefore, the set of hypotheses for the "Ideal Decision Maker" may be very large. Human Decision Makers, on the other hand, have non-negligible costs for adding low-

probability hypotheses to their current hypothesis set; costs of increasing their information-processing burden, and costs of collecting information necessary to evaluate these additional hypotheses. Finally, and most importantly, the human Decision Maker must retrieve hypotheses from memory in order to evaluate them. These considerations suggest that a human Decision Maker will generate a less-than-exhaustive hypothesis set. Next to be discussed are our previous research results (Gettys and Fisher, in preparation) bearing on the plausibility assessment model which have identified several heuristic decision rules used in plausibility assessment.

#### 2.4 Past Research on the Plausibility Model.

Our model of human plausibility assessment assumes that Decision Makers base their decisions on whether or not to include a hypothesis in the current hypothesis set on its subjective plausibility. It further assumes that the memory retrieval process is controlled by a second subjective plausibility, namely the plausibility of the entire current hypothesis set. The quantity  $P(H_1|D)$  is the normative counterpart of the subjective plausibility of a single hypothesis, and  $P(C|D)$  has the same relationship to the subjective plausibility of the entire current hypothesis set. The Gettys-Fisher study primarily attempted to discover how subjects utilize their subjective plausibilities to 1) decide if a hypothesis should be included in the current hypothesis set and 2) to determine if the hypothesis retrieval process should be initiated or terminated.

Subjects in the Gettys-Fisher experiment worked three hypothesis generation tasks, one of which has been previously discussed. Subjects estimated posterior odds for all hypotheses that they had generated. The magnitude of posterior odds for new hypotheses on the trial when they were introduced was found to be related to the odds of the most probable of the old hypotheses. The results suggested that subjects employ a heuristic rule of adding a hypothesis to the current hypothesis set only if its plausibility is high enough to make it a strong contender. In fact, 90% of the hypotheses introduced were at least half as likely as the most plausible hypothesis, and the

modal new hypothesis was either evaluated as the most plausible hypothesis, or as plausible as the most plausible hypothesis. This heuristic may have profound implications for hypothesis generation. It suggests that the basic strategy is to search for hypotheses that will be "leading contenders" in the current hypothesis set. The decision to include a new hypothesis in this set is primarily governed by what is already in the set. Presumably, the subject is comparing his candidate hypothesis with the most plausible of his current hypotheses. The subject tends to become increasingly strict in the criterion for hypothesis adoption as more hypotheses are generated; a process that works against the likelihood of obtaining an exhaustive set. However, this strategy may be rationalized by the subject as a search for a "better" hypothesis, rather than an exhaustive set. This behavior can be characterized as "solution" searching rather than an attempt to generate an exhaustive hypothesis set, as instructed.

The second major result of the Gettys-Fisher study relates to the control of the hypothesis retrieval process. New hypotheses are much more likely to be introduced when new data reduces the plausibility of the current hypotheses. This result suggests that hypothesis generation is cyclic, and usually occurs when the Decision Maker realizes that his current hypothesis set is inadequate in the light of the new data. This process is a second major heuristic which we propose and is consistent with much of the concept identification literature which suggests that subjects engage in "win-stay, lose-shift" strategies (Kintsch, 1970) where they retain hypotheses which are consistent with the data and test new hypotheses if the old hypotheses are inconsistent with the data.

Exactly how the plausibility assessment process works is yet to be established. At present our thinking on this topic is tentative. We assume that the consistency-checking process previously described is a preliminary plausibility-screening process rather than a full-blown plausibility assessment. In many ways the boundaries between consistency checking, plausibility assessment, and probabilistic inference are arbitrary, and we now believe that many of the same mechanisms contribute to each of these processes. We believe

that each of these three processes involves retrieval of information from memory and inductive and deductive reasoning. However, these processes differ in both their goals and in the relative contributions of memory and reasoning. Perhaps the following definitions will clarify these distinctions.

1) Consistency checking may be used by the Decision Maker in his initial screening of hypotheses particularly if the data are numerous. If a hypothesis is retrieved before all data have been processed, the consistency of that hypothesis will be checked with the remaining data. The hypothesis will be abandoned if inconsistent information is retrieved from memory. This process operates very rapidly and involves only superficial memory searching and reasoning. Its major purpose is to screen hypotheses for obvious defects before subjecting them to a more exhaustive analysis.

2) Plausibility assessment involves more reflection and deeper analysis. The major goals are to decide if the hypothesis is sufficiently plausible to warrant including it in the current hypothesis set and to decide if the current hypothesis set is sufficiently exhaustive. These goals are achieved by reasoning with facts and information retrieved from memory using the data and the hypothesis as retrieval cues. The concept of availability proposed by Tversky and Kahneman (1973, 1974) with its emphasis on frequency and ease of memory retrieval is important here. It may be that the availability of a hypothesis determines its plausibility assessment. Additionally, the plausibility of the current hypotheses set may be judged by a "metamemory" process (Lindsay and Norman, 1977); metamemory is the information we have about the contents of our memory store. For example, if you are asked to give the data of George Washington's inauguration, your metamemory gives you an indication of the likelihood of retrieving this data. Plausibility assessment of the current hypothesis set may be based on metamemory, it may be based on the ease of hypothesis retrieval or the number of hypotheses retrieved, or it may be based on all of these processes. Perhaps if your metamemory suggests that you are knowledgeable in the area, and if your memory searches no longer successfully retrieve hypotheses, then you conclude that you have retrieved all of the relevant hypotheses. This conclusion would lead to a high plausibility assessment for the current hypothesis set.

We include these ideas on the mechanism of plausibility assessments, speculative as they are, to document some of our current interests. They will not

become a formal part of our model until supported by more data. At this time, all we can say is that the goal of plausibility assessment is to determine if hypotheses or groups of hypotheses are plausible enough to process further. We have identified several heuristic strategies which may be employed by the Decision Maker in plausibility assessment.

3) Probabilistic inference concentrates on the assignment of subjective probability measures to specified, enumerated outcomes. Usually the emphasis is on identifying the most probable hypotheses, or hypothesis. It also involves reasoning with information retrieved from memory but differs from plausibility assessment in that the hypotheses are accepted as given; no attempt is made to generate new hypotheses.

Our current model for the plausibility assessment process is presented in figure 3. The input to this process are hypotheses which have been retrieved from memory and may have been checked for consistency. If the hypotheses have survived a consistency-checking process it can be assumed that they are at least minimally plausible, in the sense that no data inconsistent with their possibility has been found.

-----  
INSERT FIGURE 3 ABOUT HERE  
-----

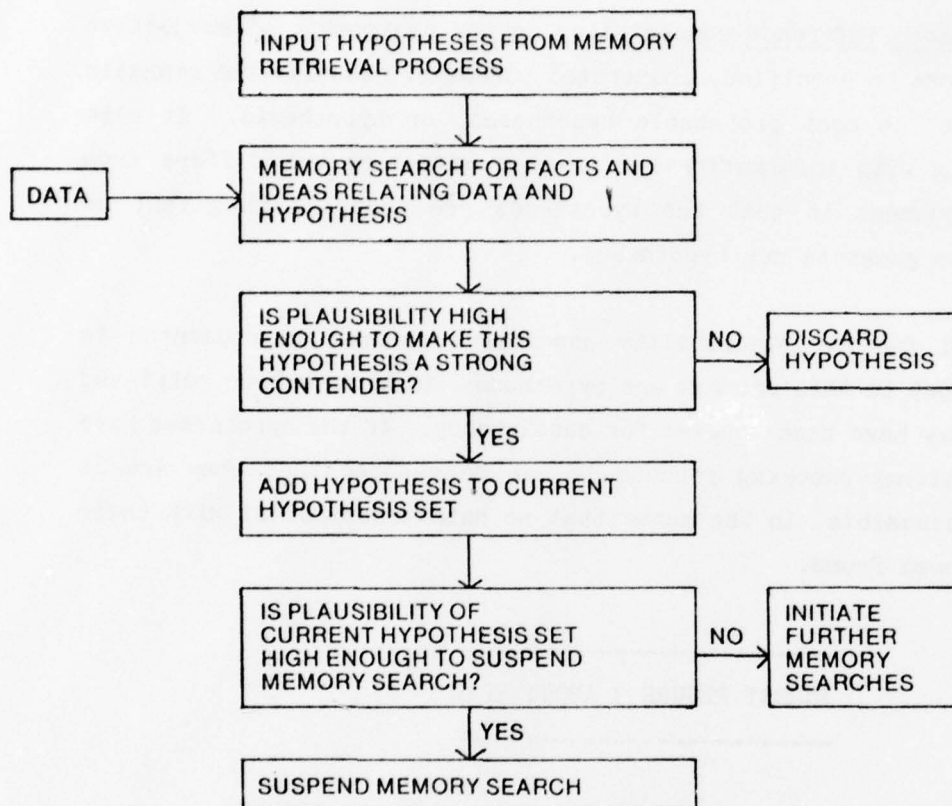


Figure 3. A model for the plausibility assessment process.

### 3.0 THE MEMORY RETRIEVAL EXPERIMENT

Most memory theorists have been concerned with retrieval from memory using a single retrieval cue. Hypothesis retrieval, however, typically involves the retrieval from memory using many data as retrieval cues, and therefore must have an additional mechanism for retrieving hypotheses which are consistent with most or all of the data. Consequently, an assumption that memory is searched using each datum in turn, and that all hypotheses retrieved by any of these single-datum memory searches are processed further, does not usually produce hypotheses that are consistent with all or most of the data. In fact, the number of candidate hypotheses produced by such a process would be unmanageable in size and the number of relevant hypotheses would be only a small percentage of the total.

However, making the opposite assumption, that a hypothesis must be retrieved for all data is equally unsatisfactory because far too few hypotheses would be produced by such a process. Successful retrieval from memory requires the presence of the necessary information in memory, and its accessibility. As the number of data increase the probability of retrieving any particular hypothesis for all data becomes vanishingly small.

These two assumptions do have the virtue that they are limiting-cases and define the ends of a continuum along which human performance must necessarily fall. The basic strategy employed in this experiment was to develop a model which could serve as a yard-stick to measure where human performance falls on this continuum. The model then can be used as a measuring device where its free parameter estimates this location. This experimental strategy has been employed by Greeno (1970) who has applied the idea of using a model as a measurement tool to questions in human learning.

#### 3.1 Details of the Memory Tagging Model

The model assumes that hypotheses may be tagged in response to datum. In the case of a multi-data hypothesis search involving  $i$  data, the number of tags,  $X_i$ , is a random variable,  $0 \leq X_i \leq I$ . It is also assumed that the subjects

employ a variable response criterion  $C$ , which is adjusted by the subject depending on the number of data, the instructions, and the amounts of task-relevant information in memory. For these reasons it seems appropriate to assume that  $C$  is also a random variable,  $1 \leq C \leq I$ . We assume that  $C$  is distributed in the interval 1 to  $I$  with probabilities obtained from the binomial distribution ( $N = I - 1, P$ ). This distribution was chosen because it is discrete, single-peaked, and because it approximates the normal as  $N$  becomes large. Basically, the assumption of a variable criterion which is probabilistically distributed implies that the average criterion,  $\bar{C}$ , is the mean of a random variable. If  $X \geq C$ , where  $C$  is the criterion number of tags on that trial for that subject, then the subject responds with that hypothesis. If  $X < C$  then no response is made.

The mean criterion,  $\bar{C}$ , is the free parameter of interest. It is bounded by 1 and by  $I$ , which correspond to the two limiting cases discussed earlier. In the present experiment the basic strategy is to estimate  $\bar{C}$  from the single-datum and multi-data retrieval probabilities and the model. The estimate of  $\bar{C}$  locates human performance along the continuum of interest.

### 3.2 Method and Procedure

3.2.1 Design. Each subject performed three hypothesis retrieval tasks, a task where one datum was used as a retrieval cue and two other tasks where three and six data were used. The assignment of tasks to these three conditions was counterbalanced, so that for each task an equal number of subjects were run in the one, three, and six data conditions. Order of tasks and order of presentation were also counterbalanced. Therefore, each subject performed once in each of the tasks and once in each of the three conditions.

3.2.2 Hypothesis Generation Tasks. The three hypotheses generation tasks which were used were those previously employed by Gettys and Fisher (in preparation). Each task involved the generation of hypotheses. In the "States" task the possible hypotheses were the 50 U.S.A. States and the data were notable products and industries of one of these States. In the

"Occupations" task the hypotheses were occupations of skilled tradesmen and the data were tools typically used by a workman in one of these occupations. In the "Majors" task, various majors at the University of Oklahoma were the hypotheses, and the data were classes taken by University of Oklahoma students. In the various tasks, subjects were given one, three, or six data and were told to respond with as many hypotheses as possible which occurred to them. Subjects were told to respond with any hypothesis that came to mind after inspecting the data without being concerned with its plausibility or lack of plausibility.

These tasks were carefully chosen with the following criteria in mind. First, we wanted tasks which were within the competence of male College Freshmen. For this reason tasks involving special expertise and training seemed inadvisable. Second, we wanted tasks where our subjects would be roughly equivalent in terms of their memory store, suggesting that tasks which were based on common information which all Freshmen should possess would be preferable. Finally, the tasks should have large numbers of plausible hypotheses for the given data. The "States" task has 50 potential hypotheses. The "Occupations" task has several thousand potential hypotheses, and there are over 200 Majors at the University of Oklahoma. The data used in the three tasks is presented in Table 1.

-----  
INSERT TABLE 1 ABOUT HERE  
-----

In the six-data task all six data were presented on a single trial. The order of presentation of the data was randomized in the one-datum and six-data conditions for every subject. The order of data within each three-data cluster in the three-data condition was also randomized.

3.2.3 Subjects. 164 male University of Oklahoma introductory psychology students served as subjects in partial fulfillment of course requirements. Subjects were assigned to conditions according to a predetermined random block order, a design which called for 144 subjects. Data from the additional 20 subjects was not used due to experimental errors.

TABLE 1: DATA USED IN THE THREE MEMORY SEARCH TASKS

STATES	OCCUPATIONS	MAJORS
1. Beef	Hammer	Psychology I
2. Fish	Drill	U.S. History
3. Aerospace Industry	Saw	Industrial Psychology
4. Citrus Fruit	Wrench	Design/Measurement of Work
5. Tourists	Pipe Threader	Personnel Management
6. Cypress Trees	Blow Torch	The Behavior of Organizations

### 3.3 Procedure

3.3.1 Instructions. Upon entering the lab, subjects were seated in front of an intelligent graphics terminal which presented all instructions and stimuli, and also recorded all responses. Once the subject was given informed-consent information, instructions on how to use the terminal keyboard and to correct typing errors were presented. Then two demonstration problems were presented. These problems consisted of giving three traits and/or characteristics of animals, and naming an animal which was consistent with all traits. After the two demonstration problems, the subject was given a similar practice hypothesis generation problem where his task was to generate "animal" hypotheses. Subjects were given 60 seconds to type in their hypotheses. Typing time was not included in the 60 second interval. Once the "return" key was pushed to enter the hypothesis, it disappeared from the screen. This procedure was used to reduce the possibility that subjects would generate hypotheses which were associates of previously-generated hypotheses.

3.3.2 Data Collection. After the practice problem was completed, subjects were given the first experimental problem. The data were presented (i.e. products of States, tools, or classes) and subjects were told to type in any hypotheses which came to mind that were relevant to the data. They were also told that their responses should be based upon all the data presented on one trial. Nine trials were always presented to each subject in the experimental phase. One six-data trial involved generating hypotheses in response to all six data of one of the tasks. Two three-data trials involved generating hypotheses to the first three data and the last three data of another task. There were six single-datum trials which involved generating hypotheses for each of the individual datum of the remaining task. The order of presentation of tasks was counterbalanced, and each task was presented as either one, two, or six trials for each subject. The order of the data within each multiple-data trial was randomized for each subject. The order of the single-datum trials was also randomized. A 60 second interval was given for subjects to generate hypotheses, just as in the practice problem. The clock which measured the

37

60 second interval was stopped when the subject typed the first letter of each hypothesis, and was restarted when the carriage return key was pressed. This procedure was used to subtract typing time from memory search time because of the large individual differences among the subjects in typing speed. Under these conditions 60 seconds was ample time to respond; subjects almost always ceased hypothesis generation before the end of the 60 second interval.

### 3.4 RESULTS AND DISCUSSION

3.4.1 The calculations of the predictions of the model. The data of this experiment were the hypotheses retrieved from memory for each of the three tasks for either one, three, or six data. Each of the nine problems was attempted by 48 subjects. These results were used to calculate the probability of retrieval of various hypotheses in the various retrieval conditions. Thus, for a hypothesis of interest, nine retrieval probabilities could be estimated. These were six one-data, two three-data, and one six-data retrieval probabilities.

These retrieval probabilities were then used as inputs of the memory-tagging model; the single-datum retrieval probabilities were used to predict the three-data and six-data retrieval probabilities and to estimate  $\bar{C}$ , the average criterion for both multi-data conditions.

A more formal derivation of the predictions of the model and two proofs that the estimation of the average criterion is unique are presented in appendix A. Here a parallel explanation is provided at the conceptual level because these idea, although simple, require complicated and difficult notation if presented formally. The basic data of this study are probabilities of generating any particular hypothesis,  $P(X \geq C)$ . As mentioned previously  $X$  is the number of tags for any given hypothesis and  $C$  is the criterion employed on that trial. Therefore, both  $X$  and  $C$  are random variables:

$$X = 0, 1, \dots, I$$

and

$$C = 1, 2, \dots, I$$

The probability of generating a hypothesis is equal to

$$4) \quad P(X \geq C) = \sum_{i=1}^I P(C=i \cap X \geq i).$$

In words, a hypothesis will be generated when the number of tags,  $X$ , is greater or equal to the response criterion,  $C$ , employed on that trial. It is

assumed that the criterion is independent of the number of tags, which gives the following expression:

$$5) P(X \geq C) = \sum_{i=1}^I P(C=i)P(X \geq i).$$

The model is now written in terms of two independent processes, the response criterion process and the memory tagging process.

First to be discussed is the memory tagging process. It will be recalled that one condition of the experiment collected single-datum retrieval probabilities. In the single-datum case the number of tags,  $X$ , can either be 0 or 1. Furthermore, the only reasonable criterion that the subject can employ is  $C=1$ . Consequently, if subjects tag a hypotheses they retrieve it with probability 1.0. So we can write for this single-datum case  $P(X \geq C) = P_s(X=1)$ , where  $s$  stands for a single-datum task.

The next step is to calculate the multi-data retrieval probabilities,  $P_m(X \geq i)$  where  $m$  stands for a multi-data probability, from the single-datum probabilities,  $P_s(X=1)$ . This can be done by assuming that the probability of tagging a hypothesis in a multi-data retrieval situation for each datum is the same as the probability of tagging that same hypothesis for the same datum in a single-datum situation.

This assumption probably is at best only approximately true. We make it for mathematical tractability. It seems most plausible if memory-tagging is assumed to be a counting process, where the subject recalls the number of times that that hypothesis had been encountered in previous searches of memory.

Given this assumption, the calculation of  $P_m(X \geq i)$  is straight-forward. Consider a three-data retrieval condition from the States task where the data were 1) Citrus Fruit, 2) Tourists, and 3) Cypress Trees and the hypothesis was California. The possible outcomes of this three-datum retrieval task are shown in figure 4 in the form of a tree diagram. The probabilities shown on the tree branches are the actual single-datum retrieval probabilities  $P_s(X=1)$

and  $P_s(X=0)$ . Memory tagging is denoted by  $X$  and not tagging for that datum is denoted by  $\bar{X}$ .

-----  
INSERT FIGURE 4 ABOUT HERE  
-----

The path probabilities and the number of tags resulting are shown to the right of the tree. Shown below the tree are the calculations of  $P_m(X_{\geq i})$  which completes the calculation of the memory tagging probabilities. The calculations of  $P_m(X_{\geq i})$  for the six data case follow the same procedure, except that the tree has six data.

The calculation of  $P(C=i)$  depends on  $P$ , the binomial generating probability. An expression for  $P(C=1)$  is written as a function of  $P$  and the number of data in the task,  $I$ :

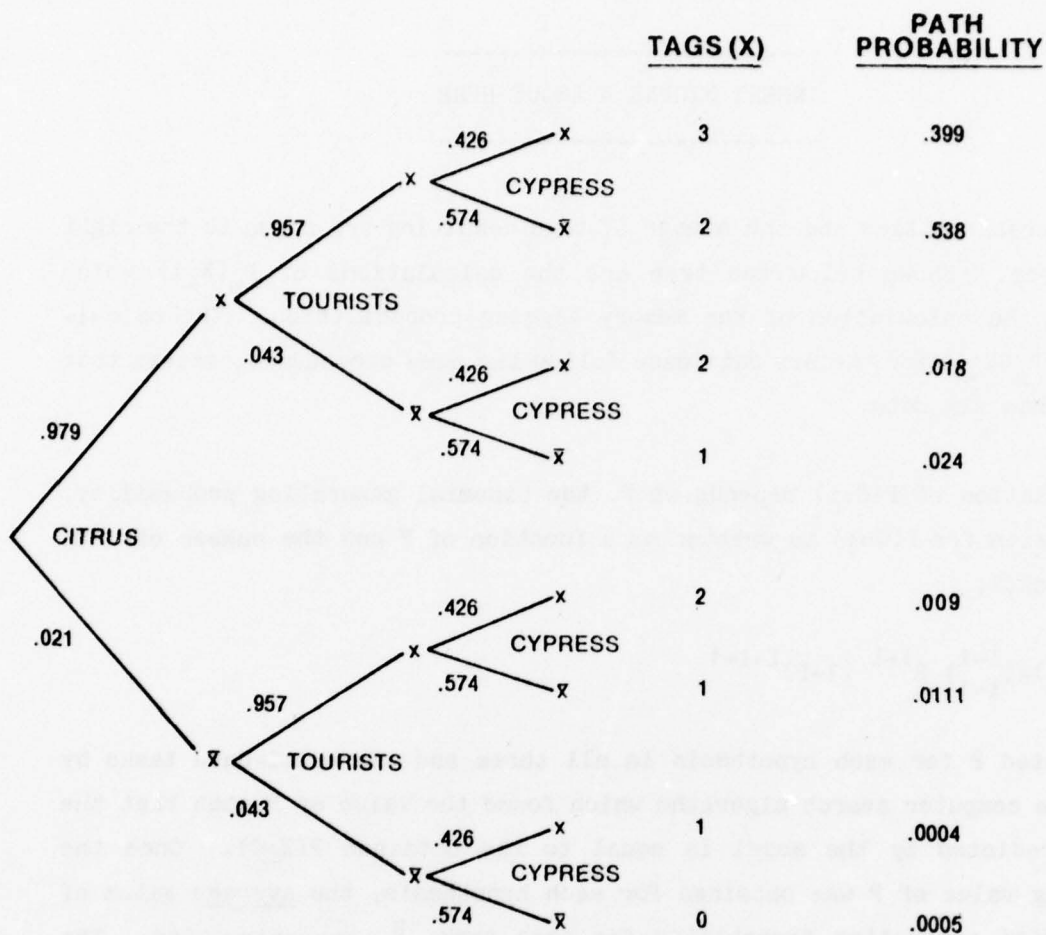
$$6) \quad P(C=i) = \binom{I-1}{i-1} P^{i-1} (1-P)^{I-i-1}$$

We estimated  $P$  for each hypothesis in all three and six multi-data tasks by means of a computer search algorithm which found the value of  $P$  such that the  $P(X_{\geq C})$  predicted by the model is equal to the obtained  $P(X_{\geq C})$ . Once the generating value of  $P$  was obtained for each hypothesis, the average value of the binomial generating probability for each task,  $\bar{P}$ , was calculated. The value of  $\bar{P}$  was then entered into an expression for  $\bar{C}$ , the average criterion value:

$$7) \quad \bar{C} = (I-1)\bar{P} + 1.$$

The estimate of  $\bar{C}$  is the desired result of the experiment, and is the average criterion number of memory tags used by the subject in the various conditions of the experiment.

If the model is to be used as a measuring device to locate the subjects along the continuum of the criterion number of tags, it is necessary to evaluate the goodness of fit of the model. This can be done by using the model and



$$P_m(X \geq 3) = .399$$

$$P_m(X \geq 2) = .399 + .538 + .018 + .009 = .964$$

$$P_m(X \geq 1) = .964 + .024 + .0111 + .0004 = .9995$$

$$P_m(X \geq 0) = .9995 + .0005 = 1.000$$

Figure 4. The calculation of  $P(X \geq 1)$  from the single - datum retrieval probabilities.

estimating  $P_m(X \geq C)$  for each hypothesis in the task using the  $\bar{P}$  as the binomial generating probability. (This is equivalent to using  $\bar{C}$  and gives the same result.) Here, rather than estimating  $\bar{P}$  by use of the model,  $P(X \geq C)$  is estimated from the model and  $\bar{P}$ .

If the model fits the data well we would expect to find a high linear relation between the predicted and the obtained  $P(X \geq C)$  values, and the slope of the best-fitting line between the obtained and predicted  $P(X \geq C)$  values should be 1.0.

Table 2 gives the  $\bar{C}$  estimates which were made for all hypotheses that had non-zero  $P_s(X \geq C)$  and  $P_m(X \geq C)$  retrieval probabilities for all trials. The number of hypotheses meeting this criteria is shown for each condition. A list of these hypotheses and their predicted and obtained retrieval probabilities is given in appendix B. Also shown in table 2 are various goodness-of-fit indices.

The results of primary interest are the  $\bar{C}$  estimates. As can be seen in table 2 the  $\bar{C}$  estimates for the six three-data problems range between 1.89 and 2.23. The average value across all three-data problems of  $\bar{C}$  is 2.01. The  $\bar{C}$  estimates for the six-data problems are higher, ranging from 2.42 to 3.20, and are on the average equal to 2.88.

-----  
INSERT TABLE 2 ABOUT HERE  
-----

The yardstick that the model affords suggests that about 2.0 memory tags are necessary for hypothesis generation in the three-data tasks, while about 2.9 tags are necessary in the six-data tasks. The rate of increase of  $\bar{C}$  as a function of the number of data in the task is interesting. If we assume that  $\bar{C}=1$  for one-datum tasks, then we have  $\bar{C}=1$ ,  $\bar{C}=2.01$ , and  $\bar{C}=2.88$  for the one, three, and six-data tasks, respectively. The value of  $\bar{C}$  used increases much more slowly than the number of data.

The major conclusion supported by these results is that hypotheses are generated in multi-data retrieval tasks when they are tagged by two or three data.

TABLE 2

ESTIMATION OF THE CRITERION NUMBER OF MEMORY TAGS,  $\bar{c}$ , AND  
GOODNESS-OF-FIT INDICES

<u>3 DATA TASKS</u>	<u>HYPOTHESES</u>	<u><math>\bar{c}</math></u>	<u>r</u>	<u>b</u>	UPPER AND LOWER 95% CONFIDENCE INTERVALS FOR b.	
					<u>LOWER</u>	<u>UPPER</u>
States $d_1$ - $d_3$	15	1.96	.935	1.144	.885	1.403
States $d_4$ - $d_6$	14	1.92	.962	1.997	.818	1.176
Majors $d_1$ - $d_3$	20	2.23	.887	1.352*	1.004	1.701
Majors $d_4$ - $d_6$	16	1.89	.892	1.459*	1.035	1.882
Occupations $d_1$ - $d_3$	17	1.98	.947	1.099	.894	1.305
Occupations $d_4$ - $d_6$	14	<u>2.10</u>	<u>.872</u>	<u>1.138</u>	.735	1.540
MEANS:		2.01	.923	1.198		

\*p &lt; .05

6 DATA TASKS

States $d_1$ - $d_6$	9	3.20	.987	.9540	.821	1.087
Majors $d_1$ - $d_6$	12	3.04	.923	1.183	.835	1.531
Occupations $d_1$ - $d_6$	8	<u>2.42</u>	<u>.874</u>	<u>1.030</u>	.457	1.603
MEANS:		2.88	.949	1.056		

This result has located human performance along a continuum of possible models. Clearly the two limiting-case models discussed previously can be rejected on the basis of these results. Any future hypothesis generation models that are developed should have the characteristic that hypotheses are retrieved from part, but not all of the data.

3.4.2 The goodness of fit of the model. We have also developed a model which, despite its simplicity, fits the data well. The coefficients of correlation,  $r$ , between the obtained  $P(X \geq C)$  and the predicted model are shown in the middle of table 2.

The model fits the data about as well as could be expected. The mean correlations are .923 and .949 calculated from the Fisher  $r$  to  $z$  transform for the three-data and six data tasks respectively. The  $P_m(X \geq C)$  probabilities are unreliable because they are estimated from 48 trials. For 48 trials with  $P=.5$ , for example, the 95% confidence interval of a binomial probability is  $\pm .139$ . A calculation was performed to investigate the impact of the unreliability of the  $P_m(X \geq C)$  probability estimates. The variance of each  $P_m(X \geq C)$  estimate was calculated for each problem for a binomial process where  $P=P_m(X \geq C)$ ,  $N=48$ . The mean of these problem variances was .00230, which is an estimate of how well a perfect model would do if the only source of error was binomial variation. The variance between the model's  $P(X \geq C)$  and the obtained  $P(X \geq C)$  contains both errors of prediction of the model, and binomial variance. These variance estimates are .00828 for the three-data problems, and .00486 for the six data problem. This analysis shows that between roughly 25% to 50% of the unexplained variance is due to variation in the criterion, and hence is intrinsic variability. These analyses show that the model fits about as well as can be expected, and that it predicts nearly all the variance that can be accounted for by any model.

We also calculated the slope,  $b$ , of the best-fitting line between observed  $P(X \geq C)$  and the predicted  $P(X \geq C)$ . These slopes are shown in table 2 with the upper and lower boundaries of the 95% confidence interval for these slopes. Seven of the nine slopes include 1.00 in their 95% confidence interval. However, there appears to be a tendency for the model to underestimate low  $P(X \geq C)$  and to overestimate high  $P(X \geq C)$  values.

In summary, the model is adequate for the measurement purposes for which it was designed, and fits the data considerably better than we expected. It allows us to estimate that hypothesis generation requires two memory tags in the three-data case and 2.9 tags in the six data case.

3.4.3 The minimally-adequate hypothesis set. One question of considerable theoretical and applied significance that is still to be addressed is the adequacy of the subjects' performance. Their task was to generate hypotheses. How well did they do at the hypothesis generation tasks? We have invented the concept of a "minimally-adequate hypothesis set" to characterize the adequacy of their performance. A minimally-adequate hypothesis set is defined as a set of hypotheses that is quite likely to contain the correct hypotheses. The phrase "quite likely" can be defined as the user wishes. For example it can be defined as a probability of .95 that the current hypothesis set contains the generating hypothesis. In this case, a minimally-adequate hypothesis set is defined as a hypothesis set having a plausibility of .95 (see equation 3 in section 2.3). Alternatively the minimally-adequate hypothesis set can be defined by knowledgeable experts as the minimum set that should be considered given the data. Therefore, the definition of the hypotheses which should be contained in this set can either be defined by Bayesian techniques, by knowledgeable experts, or by any combination of the two.

The minimally-adequate hypothesis set can be useful in characterizing the adequacy of the subjects' performance. Good performance would be marked by all or most subjects achieving a minimally-adequate hypothesis set.

We have established minimally-adequate hypothesis sets for the three six-data tasks. Because we were acting as knowledgeable experts, we deliberately chose these sets using a conservative criteria which was biased in favor of our subjects. There were three hypotheses in each set and these choices were based on additional library research on the tasks, and our initial knowledge when we designed the tasks. We compared the six-data hypothesis sets generated by the subjects to the minimally-adequate hypothesis set. We then counted the numbers of hypotheses common to both sets, and converted these numbers to cumulative percentages. Table 3 shows these results and the minimally-adequate hypothesis sets.

-----  
INSERT TABLE 3 HERE  
-----

As can be seen from an inspection of table 3, the percentage of subjects who achieved what might be charitably called "minimally-adequate" performance ranged between 0.0% and 36.5%. Even when the criterion of performance is relaxed still further to two of the three hypotheses, the subjects perform poorly using this extremely liberal criterion of adequate performance! If these tasks were particularly difficult, or required particular expertise, perhaps this level of performance would be satisfactory, but these tasks were deliberately chosen because their content should be familiar to our subjects.

These results, if they can be generalized to experts who are working with their specialities, support our contention that decision-aiding in hypothesis generation should be explored as a possible means of increasing the probability that the subject attains a minimally-adequate hypothesis set.

47

TABLE 3

THE CUMULATIVE PERCENTAGE OF SUBJECTS WHOSE HYPOTHESIS SETS  
INCLUDED EITHER THREE, TWO or ONE HYPOTHESES FROM THE  
MINIMALLY-ADEQUATE HYPOTHESIS SET

<u>TASK</u>	<u>MINIMALLY-ADEQUATE HYPOTHESIS SET</u>	PERCENT OF SUBJECTS RESPONDING WITH AT LEAST:		
		<u>3</u>	<u>2</u>	<u>1</u>
STATES	CALIFORNIA	36.5	66.3	100
	FLORIDA			
	TEXAS			
OCCUPATIONS	PLUMBER	4.3	47.8	91.3
	ELECTRICIAN			
	OILFIELD WORKER			
MAJORS	PSYCHOLOGY	0.0	36.0	84.0
	MANAGEMENT			
	INDUSTRIAL ENGINEERING			

#### 4.0 THE VERIDICAL PLAUSIBILITY STUDY

##### 4.1 Purpose

Gettys and Fisher (in preparation) demonstrated that the subjective plausibility of new hypotheses influences whether or not these hypotheses will be employed in an inference task, and that the plausibility of the current hypothesis set partially controls the memory search process. However, this study did not examine the accuracy of the subjective plausibility reports. The basic strategy of the veridical plausibility study was to obtain plausibility judgments in a situation where veridical plausibilities could be estimated.

The assessment of the accuracy of plausibility estimates is quite important in determining the most effective location of decision-aiding efforts. The memory retrieval experiment discussed previously showed that subjects are poor at retrieving minimally-adequate hypothesis sets; thus making hypothesis generation an obvious target for decision aiding. In the veridical plausibility study a similar assessment of the plausibility estimation process is made to gain some understanding of human capabilities in this aspect of hypothesis generation.

##### 4.2 Method

4.2.1 An Overview of the Method. The design of this study is based on the plausibility assessment model presented previously in section 2.3. This model proposes that there are two plausibility assessment processes, one which evaluates individual hypotheses and a second that evaluates the entire hypothesis set. Subjects in this study made estimates corresponding to the two types of plausibility assessments.

A second independent variable in this study was the plausibility of the entire hypothesis set. The plausibility of the hypothesis set was systematically varied over three values: low, medium, or high.

A third independent variable was the number of data in each judgment task. Either one, three, or six data were employed in the various tasks. This variable was manipulated to determine if number of data has an effect on plausibility judgments, as it has on Bayesian inference tasks (Slovic and Lichtenstein, 1971).

Finally, we manipulated whether or not the subjects knew the prior odds of the hypotheses. It became obvious early in the rather extensive series of pilot studies that plausibility estimates are considerably different than typical odds estimates, and we decided to determine if perhaps this effect was due to gross mis-estimation of the prior odds. The design employed 1) three levels of plausibility of the current hypothesis set and 2) three levels of the number of data as completely-crossed factorial within-subject variables. The presence or absence of prior odds was a between-subjects variable.

4.2.2 Subjects. Subjects were introductory psychology students. All subjects were males; 16 were randomly assigned to the "priors" condition and an equal number to the "no priors" condition.

4.2.3 Apparatus. Part of the instructions and the data collection in the actual study were under the control of a intelligent graphics terminal, a Compucolor model 8051, manufactured by the Intelligent Systems Corporation, Norcross, GA. The computer has color graphics which can be controlled by a light pen.

4.2.4 Estimation of veridical plausibilities. The basic task of this study was for subjects to estimate the plausibility of various hypothesized majors given some information on the classes an unknown undergraduate student had taken. This task was chosen because it is possible to obtain actual frequency counts from enrollment records. The data used were for all non-transfer undergraduate students enrolled at the University of Oklahoma in the Fall of 1977. The data base consisted of 116,875 enrollment records, where each record was of a class taken by a student.

Software was developed to determine from the enrollment records the frequency of majors of students who had taken either one, three, or six specified classes. The frequencies of these majors were used as relative frequency estimates of the posterior probabilities. This technique was employed so that assumptions of conditional independence were unnecessary to evaluate the posterior probabilities. The problems chosen by this technique are shown in appendix C.

This particular inference task was chosen because it seemed well suited to the requirements of the plausibility estimation study. Those requirements were 1) that the task possess a large number of possible hypotheses, 2) that the task allow the estimation of veridical plausibilities with a minimum of assumptions, and 3) that the relationships between data and hypotheses be intuitively understandable to the subjects.

This task met these three criteria, but it should be emphasized that the task was quite difficult for the subjects. The exact relationships between classes and majors is far from obvious; a modern university has multiple requirements for graduation both at the Departmental, College and University level. A student therefore chooses his program of courses partially on the basis of requirements, partially on the basis of his interests, and partially on the basis of his career goals. While the students of a university should have a good intuitive understanding of these variables in a general sense, no one person is privy to all of this information.

4.2.5 Problems. Eighteen problems having one, three or six data were presented in random order; each contained three specified hypotheses about the possible major of the unknown undergraduate student and a fourth "catch-all" hypothesis. The plausibility of the "catch-all" for a third of the problems was in the range of 0 to 33%, a third in 34 to 66% and a third in 67 to 100%. By so doing, the plausibility of the "catch-all" was crossed with number of data, with two problems nested in each cell.

Subjects assessed plausibilities using a magnitude estimation procedure; responses may be interpreted as posterior odds estimates.

4.2.6 Instructions. The "instructions" phase of the session lasted for about 20 minutes and was computer-assisted. The entire session lasted about an hour, although there was some variation since the study was subject-paced.

The instructions consisted of a graduated series of four tasks designed to familiarize the subjects with the procedures necessary to undertake the experimental task. The first task was practice in using a light pen in conjunction with the display. The second task introduced the magnitude estimation procedure to subjects. In this task the length of lines was adjusted with the light pen to estimate the relative areas of four rectangles. The third task was practice in magnitude estimation on a probabilistic inference task which involved predicting election outcomes. The last instructional task was a practice problem similar to those used in the actual experiment. Immediately following the instructional session, the data-collection phase began.

4.2.7 The Data Collection Procedure. Each of the 18 problems were presented in two frames on the computer screen. (A frame on the computer screen had a blue background and was rectangular; 30.8 cm wide by 27.5 cm high.)

- Frame One. In frame one of each trial, the subjects evaluated the three specifically-named hypotheses separately and also evaluated the "catch-all" hypothesis. This task differed little from a typical Bayesian inference task except that the catch-all estimate was employed.

Figure 5 is a color Xerox reproduction of a photograph of this display. It included four principle components. At the top there was a brief synopsis of the instructions as a constant reminder to the subject. Data and hypothesis areas displayed the data and hypotheses, respectively. In the center right of the display were four horizontal lines which the subjects adjusted using the light pen to make their plausibility estimates. When the light pen was aimed at a point on a line and was triggered, the segment left of the pen turned red. The subject was allowed to interactively adjust these lines until he was satisfied with his plausibility responses.

Figure 5. Frame one of the display as it was seen by the  
"Priors" subjects.

MATH 3703 ELEMENTARY STATISTICS

SCIENCE

ZOOLOGY

PSYCHOLOGY

ACCOUNTING

ALL OTHER MAJORS

Also shown in Figure 5 is the method of indicating prior odds. The small red "tick marks" were set at the prior odds for all four lines in the priors condition only.

-----  
INSERT FIGURE 5 ABOUT HERE  
-----

- Frame two. The second frame of each problem pitted the collection of specifically-named hypotheses against the catch-all, so that subjects estimated the relative plausibilities of two collections of hypotheses. Frame two contained the same data and hypotheses that were presented in the first frame. There were two 22.4 cm lines, having the same function as the four lines of the first frame. The three specifically-named hypotheses of the first frame were printed over the top of the upper line and the phrase "all other majors" was printed over the bottom line.

Subjects adjusted the lengths of these two lines to estimate the plausibilities of the two collections of hypotheses. The two frames were essentially identical in other details and red "tick marks" were also used to indicate the prior odds for those subjects in the "priors" condition.

### 4.3 Results and Discussion

4.3.1 Plausibility Assessment of Hypothesis Sets. A number of analyses have been performed on these data to examine various major and minor questions of interest. One question of central interest is how well subjects can estimate the plausibility of collections of hypotheses. An analysis of variance was performed to examine this question on the posterior log odds of the subjects' responses. For the frame one data, posterior log odds were calculated from equation 3 by making the appropriate transformations. For frame two, where subjects were estimating the plausibility of the collection of hypotheses directly, only a log transformation was necessary. All odds estimates were written in favor of the specified hypotheses. The results of the analysis of variance showed that the main effect of pages, plausibility, and number of data were significant at the  $p < .001$  level. The effect due to prior probabilities was non-significant. Also the number of data by plausibility interaction, and the four-way interaction involving the above variables and pages of the display were significant at the  $p < .05$  level. These results will be discussed in turn.

On frame one, subjects indirectly rated the collection of specified hypotheses as being noticeably more likely than these same hypotheses were rated on frame two ( $F = 38.17$ ;  $df = 1,30$ ;  $p < .001$ ). The mean geometric odds judgments for frame one was 2.99 and was 1.956 for frame two. This significant difference in average log odds estimates is attributable to the difference between calculating the plausibility of the current hypothesis set from odds estimates on frame one and the direct estimates of frame two. If the indirect frame one estimates were perfectly consistent with the direct frame two estimates then the means for both frames should be equal. The significant difference between the means therefore indicates that the subjects were somewhat inconsistent in the two types of estimates. This effect was anticipated, and was the motivation for obtaining the frame two estimates. Equation 3, which was used to aggregate the frame one estimates, is based on the union of mutually exclusive events, and this result shows that such a sum typically exceeds a direct estimate of the

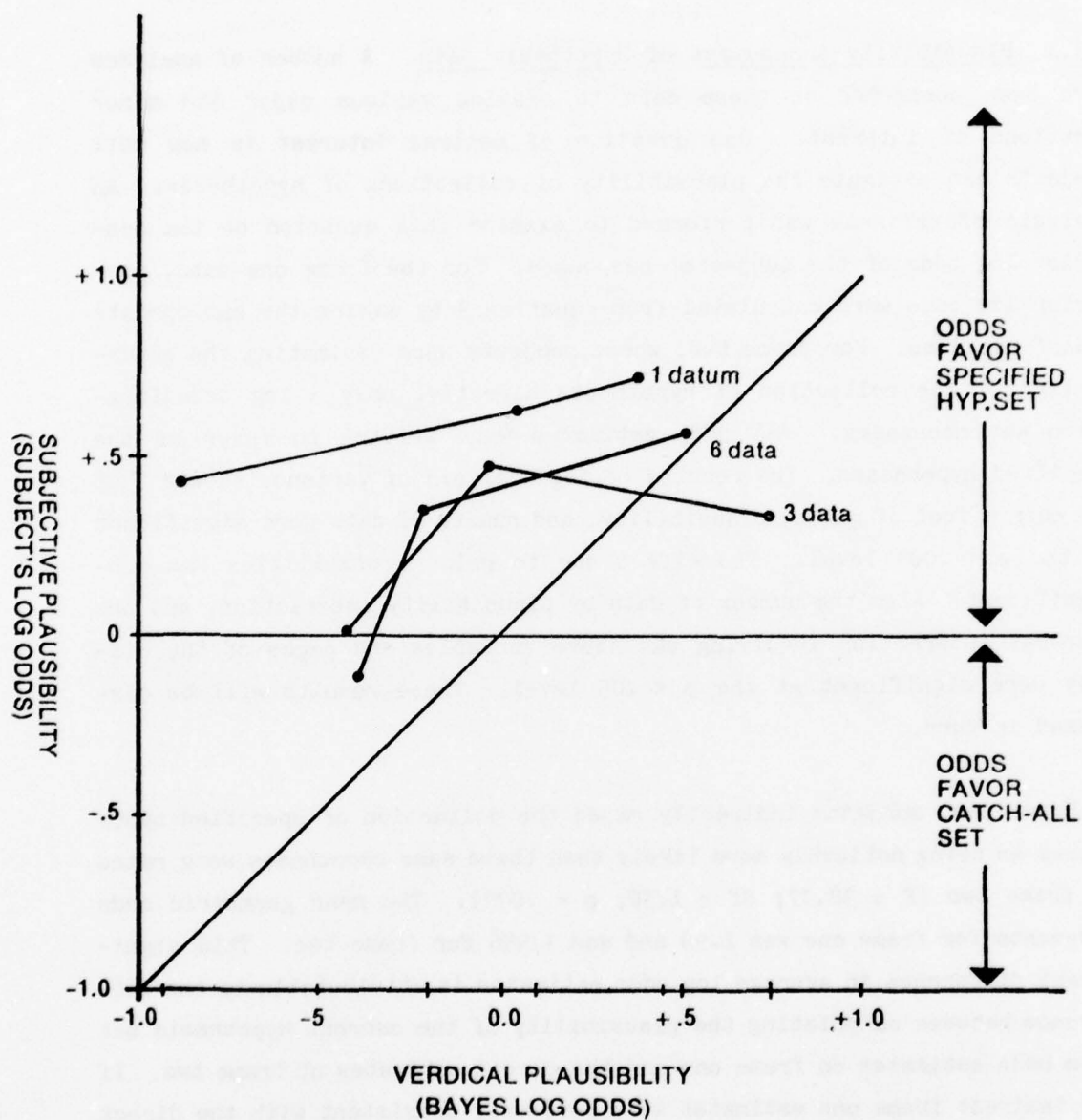


Figure 6. Subjective plausibility estimates as a function of veridical plausibility for either one, three or six data.

same union. Evidently the combination rule used by the subjects is not perfectly isomorphic with equation 3.

The effects due to number of data ( $F = 15.6$ ;  $df = 2,60$ ;  $p < .001$ ) and diagnosticity ( $F = 18.74$ ;  $df = 2,60$ ;  $p < .01$ ) were significant, as was the number of data by diagnosticity interaction ( $F = 3.51$ ;  $df = 4,120$ ;  $p < .01$ ). Figure 6 shows these results as a plot of the subjective plausibilities in log odds form versus the veridical plausibilities in log odds form. All odds ratios are written in favor of the specified hypothesis set so log values that are greater than 1.0 mean that the data favor the specified hypothesis set, while negative log values mean that the data favor the catch-all hypothesis set.

-----  
INSERT FIGURE 6 ABOUT HERE  
-----

The general form of these functions are remarkably the same in all levels of priors and no priors and frame one and frame two estimates. Perhaps the most striking aspects of these results are that the specified hypothesis set is judged more likely than the catch-all set in the vast majority of cases. Each point on this graph is the mean response to two problems. When the mean plausibility responses to the 18 problems were individually examined, the specified hypotheses are favored by 16 of the 18 mean responses. A subject who responds in a veridical fashion should favor the specified hypotheses for 9 of the 18 problems. Although the data strongly support the catch-all hypothesis set, the subjects usually still evaluate the specified hypothesis set as the more likely. Furthermore, the magnitude of their responses as compared to those of veridical subject suggests that they have difficulties in evaluating the plausibility of hypothesis sets.

Our first reaction to these results was surprise. However, when considering these results in combination with the results of the memory search experiment an explanation emerges.

First, assume that the estimation of both the plausibility of single hypotheses, and of sets of hypotheses is based on the availability of information in memory (Tversky and Kahneman, 1973). This information relates the data and hypotheses in memory. It seems logical to assume that subjects can retrieve this information in the case of specified hypotheses, since both data and hypotheses are available to act as retrieval cues. However, in the case of the unspecified hypotheses in the catch-all hypothesis set, possible hypotheses must first be retrieved from memory before the plausibility of the catch-all set can be evaluated. The deficiencies in this hypothesis retrieval process were demonstrated in the memory search experiment, so we can expect that the catch-all set will be drastically underpopulated. Consequently, the plausibility of the catch-all set relative to the specified hypothesis set will be underestimated because not enough plausible hypotheses are retrieved from memory for the catch-all set. When the plausibility of the catch-all set is underestimated, the specified hypothesis set is necessarily overestimated due to the ratio form of the odds response. The results are consistent with this explanation.

Difficulties in making plausibility estimates seem to be most pronounced in the one-datum condition. Here the subject's responses are significantly more extreme than those in the three or six-datum conditions ( $F = 32.1$ ;  $df = 1,60$ ;  $p < .001$ ), while the difference between the multi-data conditions is non-significant. This effect is probably due to the constraints we operated under in creating the multi-data problems. In order to maximize the number of students who had taken three or six classes to obtain reliable relative frequency estimates, we were forced to use popular lower-level introductory courses in most cases. While the majority of courses in the multi-data problems were lower-level, the majority of courses in the one-datum problems were upper-level. Evidently, the subjects tended to believe that the upper-level courses were more diagnostic than they actually were, thus producing this result.

The significant 4-way interaction was plotted and studied but does not seem particularly interesting. It accounts for only 5% of the variance in the study. The functions relating plausibility and number of data are quite similar regardless of values of priors and frames, and the minor changes that do occur do not change the general interpretation of figure 6.

4.3.2 Plausibility assessment of individual hypotheses. A series of analyses has been performed to examine how well subjects estimated the plausibility of individual hypotheses. In many ways these results are similar to those obtained from the previously-discussed plausibility judgments of entire hypothesis sets.

The first analysis simply counted the number of conservative and excessive judgments for all specified hypothesis and all catch-all hypotheses. This analysis was based on the median responses to each problem. As there were three specified hypotheses in each problem and 18 problems, there were 54 median judgments of specified hypotheses, and 18 judgments of catch-all hypotheses. These results are presented for the "priors" and "no-priors" conditions in table 4.

-----  
INSERT TABLE 4 ABOUT HERE  
-----

The majority of plausibility estimates of specified hypotheses were excessive, while the majority of the plausibility estimates for the catch-all hypotheses were conservative. This result is what would be expected given the previous data and the explanation for these excessive plausibility judgments.

We also converted the subject's odd estimates to plausibilities by normalizing their estimates for frame one. We then calculated the median response. This median subjective response was plotted vs. the veridical values and a line of best fit was calculated, as is often done in a Bayesian analysis. The slope of the line of best fit is a measure of calibration of the subject. If

TABLE 4

PROPORTIONS OF EXCESSIVE AND CONSERVATIVE  
JUDGMENTS IN FRAME ONE

<u>PRIORS FURNISHED</u>	EXCESSIVE	CONSERVATIVE	N
SPECIFIED HYPOTHESES	.722	.278	54
CATCH-ALL	.167	.833	18
<u>NO PRIORS</u>			
SPECIFIED HYPOTHESES	.815	.185	54
CATCH-ALL	.167	.833	18

the slope,  $b$ , is equal to 1.0 the subjects are well calibrated in a Bayesian sense. Slopes greater than one are characterized as "excessive", while slopes less than one are called "conservative". We tested the null hypothesis that the slopes are equal to zero, and also calculated a coefficient of correlation to characterize the variability or "noise" in the estimates. The results of these analyses for frame one are displayed in Table 5. One difference between these results and the results previously discussed is knowledge of the prior probabilities. Contrary to the thrust of Kahneman and Tversky's (1973) results, these subjects do appear to be influenced by knowledge of prior probabilities in estimating the plausibilities of individual hypotheses, while the effect of this variable on their plausibility estimates of groups of hypotheses is negligible. The greater slopes for the priors condition regression lines shown in Table 5 suggests that subjects are better calibrated (i.e., less conservative) when prior probability information is provided. A second powerful effect is seen in the differences in slope between specifically-named hypotheses and the catch-all. Specifically-named hypotheses are judged much more plausible than the catch-all. In the "no priors" condition the slopes differ by a factor of three. Graphically, this difference is even more striking. The two distributions of plotted points have very little overlap. We believe that this effect is due to the relatively greater "availability" (Tversky and Kahneman, 1973, 1974) of the specifically-named hypotheses vs the catch-all hypothesis as previously discussed. Evidently, the specifically-named hypotheses are available in memory, while the hypotheses in the catch-all are not readily available.

One aspect of these results that may be confusing to some is how the majority of the plausibility estimates can be excessive, while the slope of the best-fitting line is less than one. The values of the y-intercept are greater than zero, and most judgments lie above the positive diagonal.

-----  
INSERT TABLE 5 ABOUT HERE  
-----

The correlation of .357 for the catch-all in the no priors condition is also of interest. The square of this correlation is .127 which is the proportion

TABLE 5

RESULTS OF REGRESSION AND CORRELATIONAL ANALYSES  
OF FRAME ONE OF THE TASK

<u>PRIORS CONDITION</u>	Line of Best Fit ( $y = bx + a$ )		Test That $b = 0$		<u>P</u>	<u>Correlation</u>
	<u>b</u>	<u>a</u>	<u>t</u>	<u>df</u>		
SPECIFICALLY-NAMED HYPOTHESES	.507	.1276	7.34	52	<.001	.713
CATCH-ALL HYPOTHESIS	.298	.1437	2.10	16	<.05	.465
<u>NO PRIORS CONDITION</u>						
SPECIFICALLY-NAMED HYPOTHESES	.397	.1800	7.34	52	<.001	.713
CATCH-ALL HYPOTHESIS	.122	.1844	1.53	16	.1>p>.05	.357

TABLE 6

RESULTS OF REGRESSION AND CORRELATIONAL ANALYSIS  
ON FRAME TWO OF THE TASK

<u>PRIORS CONDITION</u>	Line of Best Fit ( $y = bx + a$ )		Test That $b = 0$		<u>p</u>	<u>Correlation</u>
	<u>b</u>	<u>a</u>	<u>t</u>	<u>df</u>		
CATCH-ALL HYPOTHESIS	.412	.1709	2.95	16	<.005	.594
<u>NO PRIORS CONDITION</u>						
CATCH-ALL HYPOTHESIS	.244	.2251	2.57	16	<.025	.540

NOTE: As there are only two judgments on frame two, the regression slope, the t, and the correlation are the same for the group of specifically-named hypotheses since the probability of the specifically-named hypotheses is necessarily one minus the probability of the catch-all.

TABLE 7

ORDINAL COMPARISONS OF THE PRIORS-NO PRIORS CONDITIONS MADE  
BY COMPARING THE ORDINAL PROPERTIES OF THE SUBJECT'S  
PLAUSIBILITY ESTIMATES WITH VERIDICAL PLAUSIBILITIES

<u>PROPORTION OF TRIALS WHERE:</u>	<u>PRIORS</u>	<u>NO PRIORS</u>	<u>p</u>
1) Correct Rank was Assigned to Most Likely Hypothesis			
Frame 1:	.526	.492	NS
Frame 2:	.613	.631	NS
2) Correct Rank was Assigned to Least-Likely Hypothesis (Frame 1)	.471	.462	NS
3) Correct Rank was Assigned to the Catch-All Hypothesis (Frame 1)	.399	.332	NS

of the catch-all variance accounted for by the veridical plausibilities. In this inference task, at least, there is little accuracy in the catch-all estimates on frame one judgments. Gettys and Fisher have shown that hypothesis generation is controlled by subjective plausibility, but evidently these feelings of plausibility for the catch-all aren't very accurate! Another interpretation of these results is that in the frame one judgments the subject is not very concerned with the plausibility of the catch-all, but rather tends to concentrate on the specifically-named hypotheses. Martin and Gettys (1969) have reported similar results. If this is the case, then the frame two catch-all results should show more effect due to the veridical plausibilities as the subject is confronted directly with the task of evaluation of two groups of hypotheses one of which is the catch-all. Table 6 presents an analysis of the frame 2 plausibility estimates. An inspection of table 6 shows the same effect as noted previously in regard to the priors-no priors manipulation; the no priors group is considerably less well-calibrated or more conservative. Here the veridical plausibilities account for 29.2% of the variability in the "no priors" estimates, which is an improvement over frame 1, but is still far from excellent performance.

-----  
INSERT TABLE 6 ABOUT HERE  
-----

4.3.3 Ordinal properties of plausibility assessments. We have conducted a series of analyses on the ordinal properties of the data. A good case could be made for the adequacy of the unaided human if he can rank-order hypotheses according to plausibility. By examining the ordinal relationships in his estimates as compared to the veridical orderings of hypotheses we can assess this component of his performance. Table 7 shows several interesting comparisons.

-----  
INSERT TABLE 7 ABOUT HERE  
-----

60

The tests of significance reported in Table 7 are chi-square statistics which examine the priors-no priors difference, all of which are non-significant. Obviously the subjects are ordering the hypotheses at a better than chance level. The chance expectation for frame one is .25 and for frame two is .50. While their performance is better than chance, it is far from perfect!

The lack of significant differences between the priors and no priors conditions is shown in Table 7. These results combined with those presented previously in Tables 5 and 6 suggest that the primary effect of the "priors" manipulation is to change the calibration of the subject. As the ordinal differences and the analysis of variance results are non-significant and as the primary priors-no priors effect on the regression line is a large change in slope,  $b$ , it appears that the primary effect of knowing the priors is to elevate the regression line in a veridical direction, thus improving the calibration.

## 5.0 GENERAL SUMMARY OF BOTH EXPERIMENTS

It is perhaps useful to summarize what has been learned to date because a much clearer picture of hypothesis generation is beginning to emerge.

Previous research suggested that if new evidence reduces the plausibility of the current hypothesis set, then new hypotheses will be generated. The Decision Maker searches for hypotheses that will be "leading contenders" in comparison to those that are already being entertained. The results of this study by Gettys and Fisher, (in preparation) support the idea that subjective plausibility controls the hypothesis generation process.

The memory search and plausibility estimation studies discussed here are more concerned with the subject's hypothesis generation capabilities, and have increased our understanding of these capabilities considerably.

The memory search experiment examined the key question in multi-data hypothesis retrieval, i.e., how hypotheses which are consistent with various data are retrieved. A memory-tagging model was developed, and a consistency-checking notion was presented. These ideas were combined into a model of how hypothesis retrieval from memory occurs. This model was used to arrive at the conclusion that hypotheses are retrieved if tagged by two or three data. A second, equally important aspect of this study was the evaluation of memory retrieval performance with the minimally-adequate hypothesis set. This evaluation led to the conclusion that the hypothesis retrieval process is very inefficient, and that a Decision Maker retrieves far fewer hypotheses than he should.

The plausibility estimation study shows that plausibility estimation also is deficient, and that part of this deficiency can be traced to inadequacies in hypothesis retrieval. The results suggest that hypothesis retrieval is necessary in plausibility assessment in order to properly evaluate the catch-all hypothesis. If the Decision Maker fails to retrieve many hypotheses for the catch-all, he overestimates the plausibilities of the specified hypotheses.

Paradoxically, if far too few hypotheses are generated, and if the plausibility of these few hypotheses is overestimated, the Decision Maker is left in a very vulnerable predicament where his hypothesis generation performance is quite deficient and he is unaware of his deficiencies.

One possible remedy, of course, would be to improve hypothesis retrieval by hypothesis retrieval aiding. This would both increase the number of hypotheses that the Decision Maker entertains and simultaneously improve his plausibility assessments. We are currently evaluating this possibility.

## 6.0 REFERENCES

- Anderson, J.R., and Bower, G.H. Human Association Memory. Washington, D.C.: Winston, 1973.
- Edwards, W. Nonconservative probabilistic information processing systems. Report of DSC, ESD, Air Force System Command, USAF ESD-TR-66-404, 1966.
- Gettys, C., and Fisher, S.D. Hypothesis Plausibility and Hypothesis Generation, in preparation.
- Greeno, J.G. How associations are memorized. In D. Norman (Ed.), Models of Human Memory, New York: Academic Press, 1970.
- Kintsch, W. Learning, Memory, and Conceptual Process, New York: Wiley, 1970.
- Lindsay, P.H., and Norman, D.A. Human Information Processing, New York: Academic Press, 1977.
- Martin, D.W., and Gettys, C.F. Feedback and response mode in performing a Bayesian decision task. Journal of Applied Psychology, 1969, 82, 4-8.
- Newell, A., and Simon, H. Human Problem Solving, Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1972.
- Schiffrin, R.M. Memory Search. In D. Norman (Ed.), Models of Human Memory, New York: Academic Press, 1970.
- Schiffrin, R.M., and Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84, 127-190.
- Slovic, P., and Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.
- Smith, E.E., Shoben, E.J., and Rips, L.J. Structure and process in semantic memory: A featural model for semantic decisions. Psychological Review, 1974, 81, 214-241.
- Tulving, E. Episodic and Semantic Memory. In E. Tulving and W. Donaldson (Eds.) Organization and Memory, New York: Academic Press, 1972.
- Tversky, A., and Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 1973, 5, 207-232.
- Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 85, 1124-1131.

## Appendix A

### A Model Parameter as a Unique Root

to a Nth Degree Polynomial

Thomas Mehle and Dale Umbach

The University of Oklahoma

Running head: Unique Model Parameter

Abstract

Presented in this paper are a derivation for a variable response criterion model and a demonstration that the estimate of the criterion is unique. This estimate may be of value in analyzing psychological counting process models; an example application is discussed. The possible problem with the estimate was that its uniqueness and hence its utility was questionable. General proofs of the uniqueness of the estimate are presented in traditional probabilistic notation and alternately in decision theoretic terms.

A Model Parameter as a Unique Root  
to a Nth Degree Polynomial

Counting process models have played an important role in psychological theories; the use of such models has been common place whenever a psychological process may be profitably modeled by assuming that some underlying process gives rise to discrete events which are counted. A response is made when the count meets a criterion. Included among numerous examples in the literature are models for the quantum theory of vision (Stevens, 1972; Pirene and Marriott, 1959), the availability model of human decision behavior (Tversky and Kahneman, 1973), the accumulator model of reaction time behavior (Pachella, 1974, p. 75; Kantowitz, 1974, p. 99; Audley and Pike, 1965; LaBerge, 1962; McGill, 1962), memory retrieval models (see Ratcliff, 1978, p. 74), signal detection models, (Pike, 1973; Green and Swets, 1966, p. 136) and a frequency model for verbal discrimination learning (Eckert and Kanak, 1974, p. 583).

One approach to modeling counting processes postulates that there are variations in response criteria over trials for an individual and across individuals (Gettys, Mendoza and Nicewander, Note 1.) That is, rather than the response criteria being fixed for all trials and all individuals, the criteria is assumed to have a distribution function over a range of possible values. The following section presents an example application of a variable response criterion model in which the response criterion is assumed to be binomially distributed. An attractive feature of the binomial distribution is that although the  $p$  parameter is a root to a  $n$ th degree polynomial, it is

uniquely determined in this application. Proofs of the uniqueness of the  $p$  estimate are presented in the final section.

### An Application

In a study of the processes involved in retrieving hypotheses from memory during a decision task, Gettys, Fisher and Mehle (Note 2) obtained empirical probabilities that various hypotheses were used as responses on trials for which subjects were given varying amounts of data. The notation used here will be to let  $M_{ij}(x)$  be the probability that the  $j$ th hypothesis will be recalled from memory for at least  $x$  data given that the subject was presented a set of  $i$  data,  $0 \leq x \leq i$ . Let  $R_{ij}(x)$  be the probability that the  $j$ th hypothesis will be recalled from memory for exactly  $x$  data given that the subject was presented  $i$  data. Thus

$M_{ij}(x) = \sum_{k=x}^i R_{ij}(k)$  and  $M_{ij}(0) = \sum_{k=0}^i R_{ij}(k) = 1$ . Figure 1 is a decision tree illustrating the calculation of  $R_{ij}(x)$  and  $M_{ij}(x)$  in this application, when  $i = 3$ . Let  $C_{ij}$  be the response criterion, a random variable, for hypothesis

---

Insert Figure 1 about here

---

$j$  given that  $i$  data were presented. The criterion is used in a typical counting process model decision rule: if the number of data (out of the set of  $i$  data) for which hypothesis  $j$  is recalled is equal to or greater than  $C_{ij}$ , the

subject will give hypothesis  $j$  as a response; otherwise, hypothesis  $j$  will not be used as a response. Let  $P_j(i)$  be the probability that a subject will give the  $j$ th hypothesis as a response given a set of  $i$  data.

It should be noted that  $M_{ij}(x)$ ,  $R_{ij}(x)$ ,  $C_{ij}$  and  $P_j(i)$  are also functions of which data are presented. However, incorporating labels to indicate data associations would needlessly complicate the notational scheme since data labels are not crucial in this paper's formulations.

Gettys et al. (Note 2) employed a variable response criterion model to predict  $P_j(3)$  and  $P_j(6)$  given  $P_j(1)$ . In the one-datum condition, subjects were instructed to respond with all hypotheses recalled to allow an assumption that the response criteria on single-datum trials were always one:  $R_{ij}(1) = P_j(1)$ ,  $j = 1, 2, \dots, J$ .

Next, for a given  $i$  and  $j$ ,  $R_{ij}(x)$  was calculated for  $x = 0, 1, \dots, i$ , see Figure 1. In this application, these computations were similar to using the binomial probability mass function to calculate  $P(X = x)$  for  $i$  independent Bernoulli trials, except that the trials were not Bernoulli because the success probabilities, although known, were not constant across trials but varied considerably as a function of which data were presented. However, by assuming the trials were independent,  $R_{ij}(x)$  could be calculated in the manner illustrated in Figure 1. In other applications for which an assumption of constant success probabilities would be reasonable, calculation of  $R_{ij}(x)$  could be accomplished with less tedium by using the binomial distribution.

Gettys et al. (Note 2) made the assumption that for trials on which a given hypothesis was used as a response, the criterion number of retrievals,  $C_{ij}$ , was distributed binomially, with the following probability mass function:

$$f(\underline{c}) = \binom{i}{\underline{c}} p^{\underline{c}} (1-p)^{i-\underline{c}} \quad (1)$$

It should be noted that the  $M_{ij}(\underline{x})$ 's are monotone decreasing in  $\underline{x}$ ,  $\underline{x} = 0, 1, 2, \dots, i$ . Assuming that  $C_{ij}$  is distributed according to Eq. (1), an expression for  $P_j(i)$  is:

$$\sum_{\underline{x}=0}^i M_{ij}(\underline{x}) f(\underline{x}) = \sum_{\underline{x}=0}^i M_{ij}(\underline{x}) \binom{i}{\underline{x}} p^{\underline{x}} (1-p)^{i-\underline{x}} \quad (2)$$

By viewing Eq. (2) as a polynomial in  $p$ , Gettys et al. (Note 2) used the empirical values of  $M_{ij}(\underline{x})$  to solve for  $p$ .

#### The Uniqueness Problem

A cursory examination of Eq. (2) would indicate that for any given values of  $P_j(i)$  and  $M_{ij}(\underline{x})$ ,  $p$  need not have a unique root. In fact, by the Fundamental Theorem of Algebra, Eq. (2) must have  $i$  roots in  $\mathbb{C}$ , the complex number space. If  $p$  were not unique on  $[0, 1]$ , in particular, its usefulness as a model parameter would clearly be compromised. Following are two simple proofs of the uniqueness of  $p$ , the first in decision theoretic terms and the second in traditional probabilistic notation.

Let  $f(p) = \sum_{\underline{x}=0}^i M_{ij}(\underline{x}) \binom{i}{\underline{x}} p^{\underline{x}} (1-p)^{i-\underline{x}}$ . The first clue to the behavior

of  $f(\underline{p})$  can be obtained by examining it at the endpoints of the interval containing allowable  $\underline{p}$  values, namely  $[0, 1]$ :

$$f(0) = M_{ij}(0) = 1 \geq P_j(\underline{i}) \quad (3)$$

$$f(1) = M_{ij}(\underline{i}) \leq P_j(\underline{i}) \quad (4)$$

The inequality portion of expression (4) follows from the monotone behavior of the  $M_{ij}(\underline{x})$ 's. Since  $f(0) \geq P_j(\underline{i})$  and  $f(1) \leq P_j(\underline{i})$ , and since  $f(\underline{p})$  is continuous, there must be at least one solution to  $f(\underline{p}) = P_j(\underline{i})$  for  $\underline{p} \in [0, 1]$ . Thus all that remains to be shown is that there is at most one root.

The first proof, stated in decision theoretic notation, uses the fact that the binomial  $(\underline{n}, \underline{p})$  family with fixed  $\underline{n}$  has the monotone likelihood ratio (MLR) property but can be generalized to any family of distributions with the MLR property, e.g. any exponential family with natural parameterization.

Define

$$\phi_j(\underline{x}) = \begin{cases} 1 & \text{for } \underline{x} < \underline{j} \\ 0 & \text{for } \underline{x} > \underline{j} \end{cases} \quad (5)$$

for  $\underline{j} = 1, 2, \dots, \underline{i} + 1$ ; and for  $\underline{j} = 1, 2, \dots, \underline{i}$ , define

$$\alpha_k = M_{ij}(\underline{k} - 1) - M_{ij}(\underline{k}); \alpha_{i+1} = M_{ij}(\underline{i}). \quad (6)$$

Note that one can write  $M_{ij}(\underline{x}) = \sum_{k=1}^{i+1} \alpha_k \phi_k(\underline{x})$  for  $\underline{x} = 0, 1, \dots, \underline{i}$ . Now,

$f(\underline{p})$  may be expressed as:

71

$$\begin{aligned}
 f(p) &= \sum_{j=0}^{\underline{i}} \sum_{k=1}^{\underline{i}+1} \alpha_k \phi_k(j) \binom{\underline{i}}{j} p^j (1-p)^{\underline{i}-j} \\
 &= \sum_{k=1}^{\underline{i}+1} \alpha_k \sum_{j=0}^{\underline{i}} \phi_k(j) \binom{\underline{i}}{j} p^j (1-p)^{\underline{i}-j}.
 \end{aligned} \tag{7}$$

$H_0: p \geq 1/2$  versus  $H_1: p < 1/2$

But, by Lehman (1959, p. 68) each  $\phi_k$  is a uniformly most powerful test of the binomial parameter  $p$  with fixed  $\underline{n}$  for  $H_0: p \geq 1/2$  versus  $H_1: p < 1/2$ , and

$\sum_{j=0}^{\underline{i}} \phi_k(j) \binom{\underline{i}}{j} p^j (1-p)^{\underline{i}-j}$  is strictly decreasing in  $p$  for  $p \in [0, 1]$

for each  $\underline{j} = 1, 2, \dots, \underline{i}$ . Since the  $M_{ij}(\underline{x})$ 's are monotone in  $\underline{x}$ , each  $\alpha_k \geq 0$  for  $k = 1, 2, \dots, \underline{i} + 1$ . So  $f$  is strictly decreasing if  $\alpha_k > 0$  for some  $k = 1, 2, \dots, \underline{i}$ . But this is equivalent to:

$$0 < \sum_{j=1}^{\underline{i}} \alpha_j = M_{ij}(0) - M_{ij}(\underline{i}) = 1 - M_{ij}(\underline{i}). \tag{8}$$

All that is required now is for all  $M_{ij}(\underline{x})$ 's to not be equal. This requirement may be satisfied by observing that were all  $M_{ij}(\underline{x})$ 's equal, to some common value  $\underline{M}$ ,  $P_j(\underline{i})$  must also equal  $\underline{M}$  and any value of  $p$  would work equally well. Thus, except for the preceding degenerate case, there is at most one root of  $f(p) = P_j(\underline{i})$  in  $[0, 1]$ .

Another proof that  $f$  is strictly decreasing follows from a result in Feller (1968, p. 173) that the binomial distribution function,  $F(k; \underline{n}, p)$ , can be represented:

$$F(\underline{k}; \underline{n}, \underline{p}) = (\underline{n} - \underline{k}) \binom{\underline{n}}{\underline{k}} \int_0^{1-\underline{p}} \underline{t}^{\underline{n}-\underline{k}-1} (1-\underline{t})^{\underline{k}} d\underline{t} \quad (9)$$

Suppose  $0 \leq p_1 < p_2 \leq 1$ . Define  $F_1$  and  $F_2$  as

$$F_1(\underline{x}) = F(\underline{x}; \underline{1}, p_1) \quad (10)$$

$$F_2(\underline{x}) = F(\underline{x}; \underline{1}, p_2). \quad (11)$$

Then

$$F(p_1) = \int_{[0, \underline{1}]} M_{1j}(\underline{x}) dF_1(\underline{x}) \quad (12)$$

and

$$f(p_2) = \int_{[0, \underline{1}]} M_{1j}(\underline{x}) dF_2(\underline{x}). \quad (13)$$

Thus  $f(p_1) > f(p_2)$  if and only if

$$\int_{[0, \underline{1}]} M_{1j}(\underline{x}) d[F_1 - F_2](\underline{x}) > 0. \quad (14)$$

However, integration by parts yields:

$$\int_{[0, \underline{i}]} M_{ij}(\underline{x}) d[F_1 - F_2](\underline{x}) = M_{ij}(\underline{x}) [F_1(\underline{x}) - F_2(\underline{x})] \Big|_0^{\underline{i}^*} \quad (15)$$

$$\begin{aligned} - \int_{[0, \underline{i}]} [F_1(\underline{x}) - F_2(\underline{x})] dM_{ij}(\underline{x}) &= - \int_{[0, \underline{i}]} [F_1(\underline{x}) - F_2(\underline{x})] dM_{ij}(\underline{x}) \\ &= \sum_{\underline{k}=0}^{\underline{i}-1} [F_1(\underline{k}) - F_2(\underline{k})] [M_{ij}(\underline{k}) - M_{ij}(\underline{k}+1)]. \end{aligned}$$

Now, the monotone property implies that  $M_{ij}(\underline{k}) - M_{ij}(\underline{k}+1) \geq 0$  and Eq. (8) implies for some  $\underline{k} = 0, 1, \dots, \underline{i}-1$  that  $M_{ij}(\underline{k}) - M_{ij}(\underline{k}+1) > 0$ . Thus the proof is completed once it is established that for  $\underline{k} = 0, 1, \dots, \underline{i}-1$   $F_1(\underline{k}) > F_2(\underline{k})$ . (16)

By Eq. (15), however,  $p_2 > p_1$  implies

$$\begin{aligned} F_1(\underline{k}) &= (\underline{i} - \underline{k}) \left( \frac{1}{\underline{k}} \right) \int_0^{1-p_1} \underline{t}^{\underline{i}-\underline{k}-1} (1-\underline{t})^{\underline{k}} d\underline{t} > \\ &(\underline{i} - \underline{k}) \left( \frac{1}{\underline{k}} \right) \int_0^{1-p_2} \underline{t}^{\underline{i}-\underline{k}-1} (1-\underline{t})^{\underline{k}} d\underline{t} = F_2(\underline{k}), \end{aligned} \quad (17)$$

thus completing the proof.

Reference Notes

1. Gettys, C.F., Mendoza, J.L. and Nicewander, W.A. A generalization of Thurstone's Law of Comparative Judgement. Paper presented at the Psychonomic Science Meetings, Boston, November 1974.
2. Gettys, C.F., Fisher, S.D. and Mehle, T. Data plausibility and hypothesis generation. Manuscript in preparation, 1978.

## References

- Audley, R.J. and Pike, A.R. Some alternative models of choice. The British Journal of Mathematical and Statistical Psychology. 1965, 18, 207-225.
- Eckert, E. and Kanak, N.J. Verbal discrimination learning: a review of the acquisition, transfer and retention literature through 1972. Psychological Bulletin. 1974, 81, 582-607.
- Feller, W. An Introduction to Probability Theory and its Applications: Volume 1. New York: John Wiley and Sons, Inc., 1957.
- Green, D.M. and Swets, J.A. Signal Detection Theory and Psychophysics. New York: John Wiley and Sons, Inc., 1966.
- Kantowitz, B.H. Double stimulation. In B.H. Kantowitz (Ed.), Human Information Processing: Tutorials in Performance and Cognition. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.
- LaBerge, D. A recruitment theory of simple behavior. Psychometrika. 1962, 27, 375-396.
- Lehmann, E.L. Testing Statistical Hypotheses. New York: John Wiley and Sons, Inc., 1959.
- McGill, W.J. Random fluctuations of response rate. Psychometrika. 1962, 27, 3-17.
- Pachella, R.G. The interpretation of reaction time in information-processing research. In B.H. Kantowitz (Ed.), Human Information Processing: Tutorials in Performance and Cognition. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.

- Pike, R. Response Latency models for signal detection. Psychological Review. 1973, 80, 53-68.
- Pirenne, M.H. and Marriott, F.H.C. The quantum theory of light and the psycho-physiology of vision. In S. Koch (ed.), Psychology: A study of a Science; Study I. Conceptual and Systematic; Volume 1. Sensory, Perceptual and Physiological Formulations. New York: McGraw Hill, 1959.
- Ratcliff, R. A theory of memory retrieval. Psychological Review. 1978, 85, 59-108.
- Stevens, S.S. A neural quantum in sensory discrimination. Science. 1972, 177, 749-762.
- Tversky, A. and Kahneman, D. Availability: a heuristic for judging frequency and probability. Cognitive Psychology. 1973, 5, 207-232.

Footnote

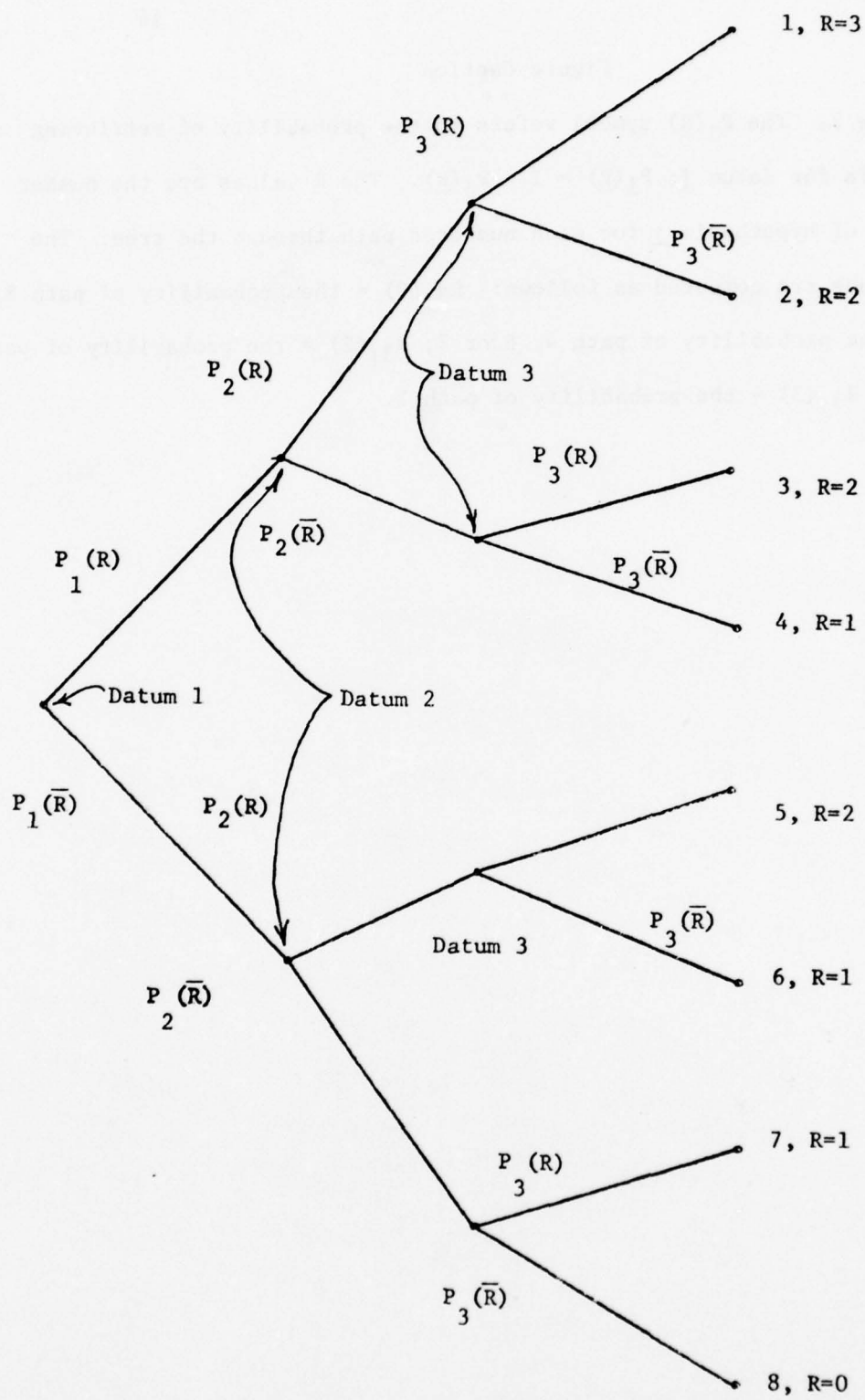
This work was partially supported by the Office of Naval Research, N00014-77-C-0615, NR 197-040.

The authors are grateful for the helpful comments of Charles Gettys, Jack Kanak, and Allen Nicewander. The second proof (in probabilistic notation) was contributed by Bradford Crain, Department of Mathematics, Portland State University

Requests for reprints should be sent to Dale Umbach, Department of Mathematics, Physical Sciences Building, the University of Oklahoma, Norman, Oklahoma 73019.

## Figure Caption

Figure 1. The  $P_j(\underline{R})$  symbol refers to the probability of retrieving a hypothesis for datum  $\underline{j}$ ;  $P_j(\overline{\underline{R}}) = 1 - P_j(\underline{R})$ . The  $\underline{R}$  values are the number of recalls of hypothesis  $\underline{j}$  for each numbered path through the tree. The  $R_{3j}(\underline{x})$  values are computed as follows:  $R_{3j}(0)$  = the probability of path 8;  $R_{3j}(1)$  = the probability of path 4, 6 or 7;  $R_{3j}(2)$  = the probability of path 2, 3 or 5;  $R_{3j}(3)$  = the probability of path 1.



## Appendix B

### Predicted versus Empirical Hypothesis Recall Probabilities

Data and Hypotheses	Predicted	Empirical
<hr/>		
Data: Beef, Fish and Aerospace Industry		
Texas	.849	.898
Louisiana	.154	.082
California	.512	.694
Florida	.550	.796
Georgia	.042	.082
Oklahoma	.433	.204
Colorado	.176	.102
New York	.106	.122
Oregon	.160	.184
Alabama	.029	.082
Hawaii	.079	.082
Missouri	.047	.020
Tennessee	.018	.020
Illinois	.054	.020
Wyoming	.126	.020
<hr/>		
Data: Citrus Fruit, Tourists, Cypress Trees		
Florida	.944	.959
Georgia	.249	.388
California	.857	.837

Data and Hypotheses	Predicted	Empirical
---------------------	-----------	-----------

Louisiana	.281	.388
Texas	.512	.510
Alabama	.135	.245
Hawaii	.311	.163
Arizona	.203	.122
New Mexico	.128	.082
Washington	.114	.082
Mississippi	.164	.163
Oregon	.045	.061
Nevada	.160	.061
Arkansas	.026	.041

---

Data: Beef, Fish, Aerospace  
Industry, Citrus Fruit, Tourists,  
Cypress Trees

Texas	.779	.750
Florida	.877	.896
California	.780	.708
Louisiana	.178	.208
Hawaii	.143	.125
Oregon	.060	.125
Georgia	.092	.083
Oklahoma	.199	.083
Alabama	.044	.083

---

Data: Psychology I, U.S. History,  
Industrial Psychology

Architecture	.020	.021
English	.095	.043
Pre Med	.083	.191
Zoology	.054	.043
Biology	.024	.021

## Data and Hypotheses

## Predicted

## Empirical

---

Pre Law	.104	.170
Sociology	.211	.191
Education	.202	.234
History	.253	.426
Mathematics	.066	.021
Chemistry	.064	.064
Business	.206	.298
Finance	.042	.085
Management	.117	.106
Economics	.059	.043
Accounting	.089	.043
Public Relations	.027	.043
Political Science	.152	.149
Engineering	.154	.149
Mechanical Engineering	.020	.021

---

Data: Design/Measurement of Work,  
Personnel Management, The Behavior  
of Organizations

Psychology	.378	.319
Sociology	.229	.255
Education	.080	.106
Political Science	.140	.064
Anthropology	.079	.043
Advertising	.039	.021
Economics	.094	.064
Business	.335	.532
Management	.352	.638
Accounting	.203	.298
Finance	.114	.170
Business Administration	.137	.149
Marketing	.121	.149

Data and Hypotheses	Predicted	Empirical
---------------------	-----------	-----------

English	.033	.021
Engineering	.184	.106
Industrial Engineering	.039	.021

Data: Psychology I, U.S. History,  
Industrial Psychology, Design/Measurement  
of Work, Personnel Management, The Behavior  
of Organizations

Psychology	.665	.755
Political Science	.164	.184
Sociology	.264	.204
Education	.159	.143
English	.064	.020
Business	.327	.388
Management	.261	.490
Accounting	.145	.122
Economics	.071	.061
Business Administration	.061	.122
Finance	.066	.082
Engineering	.189	.061

Data: Hammer, Drill, Saw

Plumber	.234	.152
Carpenter	.909	.978
Mechanic	.232	.239
Electrician	.114	.130
Machinist	.083	.108
Metal/Steel Worker	.102	.065
Welder	.102	.065
Construction	.206	.478
Artist	.027	.043
Stagehand	.021	.022

Data and Hypotheses	Predicted	Empirical
---------------------	-----------	-----------

Roofer	.054	.065
Dentist	.096	.022
Physician	.072	.022
Furniture Manufacturer	.096	.174
Framer	.049	.022
Mason	.032	.022
Farmer	.055	.043

Data: Wrench, Pipe Threader,  
Blow Torch

Plumber	.608	.804
Carpenter	.073	.109
Mechanic	.508	.304
Electrician	.140	.065
Metal/Steel Worker	.107	.087
Welder	.286	.587
Construction	.176	.197
Oil Field Worker	.128	.130
Air Conditioning Worker	.063	.043
Pipeline Worker	.162	.196
Engineer	.021	.022
Ship Builder	.030	.022
Farmer	.043	.022
Machinist	.089	.087

Data: Hammer, Drill, Saw, Wrench,  
Pipe Threader, Blow Torch

Plumber	.695	.934
Carpenter	.812	.739
Mechanic	.643	.478
Electrician	.232	.217
Machinist	.156	.130

## Data and Hypotheses

## Predicted

## Empirical

Metal/Steel Worker

.192

.152

Welder

.368

.587

Construction

.344

.261

Appendix C: Problems used in plausibility experiment  
(Probabilities are expressed as percentages)

1 datum Low catch-all plausibility

Problem 1. (Sample size = 104)

Data:	<u>Dept.-name</u> Mgt.	<u>Course-no.</u> 4363	<u>Course-title</u> Organizational Behavior
Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Management	5.7%	64.4%
	2)Political Sci.	1.9%	1.9%
	3)Law Enf. Admin.	1.1%	2.9%
	4)All others	91.3%	30.8%

Problem 2. (Sample size = 62)

Data:	<u>Dept.-name</u> PSC	<u>Course-no.</u> 4803	<u>Course-title</u> Criminal Legal Proc.
Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Political Sci.	1.9%	22.6%
	2)History	.9%	4.8%
	3)Law Enf. Admin.	1.1%	50.0%
	4)All others	96.1%	22.6%

1 datum Medium Catch-all plausibility

Problem 3. (Sample size = 152)

Data:	<u>Dept.-name</u> Zoo.	<u>Course-no.</u> 3333	<u>Course-title</u> Genetics
Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Chemistry	1.1%	9.2%
	2)Psychology	2.4%	9.2%
	3)Zoology	2.4%	34.9%
	4)All others	94.1%	46.7%

Problem 4. (Sample size = 127)

Data:	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
	PSC	3853	Prin. Criminal Inv.
Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Management	5.7%	7.1%
	2)Law Enf. Admin.	1.1%	31.5%
	3)Political Sci.	1.9%	18.1%
	4)All others	91.3%	43.3%

1 datum High catch-all plausibility

Problem 5. (Sample size = 182)

Data:	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
	Math	3703	Elementary Stat.
Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Zoology	2.4%	2.7%
	2)Psychology	2.4%	1.1%
	3)Accounting	5.7%	1.1%
	4)All others	89.5%	95.1%

Problem 6. (Sample size = 276)

Data:	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
	Zoo.	2255	Human Anatomy
Hypotheses	<u>Majors</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Physical Therapy	.3%	4.0%
	2)Phys. Ed.	1.3%	14.9%
	3)Zoology	2.4%	.7%
	4)All others	96.0%	80.4%

3 data-Low catch-all Plausibility

Problem 7. (Sample size = 97)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Educ.	2424	School in Am. Culture
	2)Math	2214	Anth. for Elem. Teachers
	3)Psy.	1113	Elements of Psych.

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Sociology	1.9%	1.0%
	2)All Education	8.1%	90.7%
	3)English	.7%	1.0%
	4)All others	89.3%	7.2%

Problem 8. (Sample size = 186)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Phil.	1203	Phil. Soc. and Relig. Morality
	2)Psy.	1113	Elements of Psy.
	3)Educ.	2424	School in Am. Culture

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)All Education	8.1%	57.0%
	2)All Business	21.0%	10.0%
	3)Recreation	.5%	3.2%
	4)All others	70.4%	29.6%

3 data-Medium catch-all plausibility

Problem 9. Sample size = 168)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Bot.	1114	Gen. Botony
	2)Chem.	1314	Gen. Chem.
	3)Chem.	3053	Organic Chem.

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Microbiology	1.1%	13.7%
	2)All Engineering	12.1%	11.3%
	3)Zoology	2.4%	16.1%
	4)All others	84.4%	58.9%

Problem 10. (Sample size = 133)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Zoo.	2094	Invert. Zoo.
	2)Zoo.	1121	Intro. Zoo. Lab.
	3)Phys.	2414	Gen. Phys: Mech,Sound,Heat

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Lab. Technology	.6%	3.8%
	2)Microbiology	1.1%	7.5%
	3)Zoology	2.4%	44.4%
	4)All others	95.9%	44.4%

3 data High catch-all plausibility

Problem 11. (Sample size = 263)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Math	1823	Calculus I
	2)Chem.	1314	Gen.Chem.
	3)Phil.	1203	Phil. Soc. and Relig.

Morality

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses	1)Zoology	2.4%	5.3%
	2)Geology	1.2%	6.8%
	3)Petrol. Engin.	2.1%	7.6%
	4)All others	94.3%	80.2%

Problem 12. (Sample size = 452)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-Title</u>
Data:	1)Zoo.	1121	Intro. Zool. Lab.
	2)Chem.	1314	Gen. Chem.
	3)Phys.	2414	Gen. Phys: Mech,Sound,Heat

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Microbiology	1.1%	9.1%
	2)Pharmacy	.8%	7.7%
	3)Zoology	2.4%	16.6%
	4)All others	95.7%	66.6%

6 data Low catch-all plausibility

Problem 13. (Sample size = 293)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Math.	2423	Calculus II
	2)Math	1823	Calculus I
	3)Chem.	1314	Gen. Chem.
	4)Chem.	3053	Organic Chem.
	5)Engr.	2514	Gen. Phys for Eng.and Sci.
	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Chemistry	1.1%	2.7%
	2)Accounting	6.9%	2.0%
	3)All engineering	12.1%	78.5%
	4)All others	79.9%	16.7%

Problem 14. (Sample size = 39)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Math	1444	Elem Func. & Coord. Geom.
	2)Math	1513	College Algebra
	3)Acct.	2133	Elem-Accounting I
	4)Econ.	3113	Intermed. Price Theory
	5)Hist.	1483	U.S. 1492 to 1865
	6)Psy.	1113	Elem. of Psych.
	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Psychology	2.4%	5.1%
	2)All Education	8.1%	7.7%
	3)All Business	21.0%	59.0%
	4)All others	68.5%	28.2%

6 data Medium catch-all plausibility

Problem 15. (Sample size = 177)

	<u>Dept.-name</u>	<u>Course-no</u>	<u>Course-title</u>
Data:	1)Econ.	2113	Prin. of Economics
	2)Hist.	1483	U.S. 1492-1865

3)Hist.	1493	U.S. 1065-present
4)PSC	1113	Gov't of U.S.
5)Psy.	1113	Elements of Psych.
6)Soc.	1113	Intro. to Sociology

Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)History	.9%	4.0%
	2)Zoology	2.4%	2.3%
	3)All Business	21.0%	35.0%
	4)All others	75.7%	58.7%

Problem 16. (Sample size = 158)

Data:	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
	1)Chem.	1614	Chem. for Non-Science
	2)Econ.	2113	Prin. of Economics
	3)Econ.	2843	Elements of Stat.
	4)Psy.	1113	Elements of Psych.
	5)Fin.	3303	Business Finance
	6)Math	1444	Elem. Func. and Coord. Geo.

Hypotheses:	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
	1)Accounting	6.9%	32.9%
	2)Economics	.2%	.6%
	3)All Education	8.1%	2.5%
	4)All others	84.8%	64.0%

6 data High catch-all plausibility

Problem 17. (Sample size = 470)

Data:	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
	1)Chem	1314	Gen. Chem.
	2)Math	1823	Calculus I
	3)Math	2423	Calculus II
	4)Phys.	2514	Gen. Phys. for Eng. & Sci.
	5)PSC	1113	Gov't. of U.S.
	6)Engr.	1112	Intro. to Engineering

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Electrical Engr.	2.1	16.0
	2)Meteorology	.7%	2.6%
	3)Accounting	6.9%	2.3%
	4)All others	90.3%	79.1%

Problem 18. (Sample size = 58)

	<u>Dept.-name</u>	<u>Course-no.</u>	<u>Course-title</u>
Data:	1)Zoo.	1114	Intro. to Zool.
	2)Zoo	1121	Intro. Zool. Lab.
	3)Phys.	2414	Gen. Phys: Mech, Sound, Heat
	4)Econ.	3113	Intermed. Price Theory
	5)Hist.	1483	U.S. 1492-1865
	6)Psy.	1113	Elements of Psych.

	<u>Major</u>	<u>P(H)</u>	<u>P(H/D)</u>
Hypotheses:	1)Phys. Ed.	1.3%	12.1
	2)Chemistry	1.1%	5.1%
	3)Zoology	2.4%	15.5%
	4)All others	95.2%	67.2%

OFFICE OF NAVAL RESEARCH, CODE 455  
TECHNICAL REPORTS DISTRIBUTION LIST

Director, Eng. Psychology

Programs, Code 455  
Office of Naval Research  
800 North Quincy Street  
Arlington, VA 22217 (5 cys)

Defense Documentation Center  
Cameron Station  
Alexandria, VA 22314 (12 cys)

Dr. Stephen J. Andriole  
Acting Dir.,  
Cybernetics Tech. Office  
Advanced Research Pro. Agency  
1400 Wilson Blvd  
Arlington, VA 22209

Cdr. Paul Chatelier  
OAD (E&LS) ODDR&E  
Pentagon, Rm 3D 129  
Washington, D.C. 20301

Director, Naval Analysis Prog.  
Code 431  
Office of Naval Research  
800 North Quincy St.  
Arlington, VA 22217

Director, Operations Research  
Programs, Code 434  
Office of Naval Research  
800 North Quincy Street  
Arlington, VA 22217

Dir., Statistic and  
Probability Prog. Code 436  
Office of Naval Research  
800 North Quincy St.  
Arlington, VA 22217

Office of the Chief of Naval  
Operations, OP987H  
Personnel Logistics Plans  
Dep. of the Navy  
Washington, D.C. 20350

Director, Information Systems

Program, Code 437  
Office of Naval Research  
800 North Quincy Street  
Arlington, VA 22217

Commanding Officer  
ONR Branch Office  
ATTN: Dr. J. Lester  
Building 114, Section D  
Summer Street  
Boston MA 02210

Commanding Officer  
ONR Branch Officer  
ATTN: Dr. Charles Davis  
536 South Clark Street  
Chicago, Il. 60605

Commanding Officer  
ONR Branch Officer  
ATTN: Dr. E. Gloye  
1030 East Green Street  
Pasadena, CA 91106

Dr. Bruce McDonald  
Office of Naval Research  
Scientific Liaison Group  
American Embassy, RmA-407  
APO San Francisco 96503

Dir., Naval Research Lab.  
Technical Inf. Division  
Code 2627  
Washington, D.C. 20375  
(6 cys)

Naval Research Laboratory  
ATTN: Code 5707  
Washington, D.C. 20375

Dr. Gary Poock  
Operations Research Dep.  
Naval Postgraduate School  
Monterey, CA 93940

CDR Paul Nelson  
Naval Medical R&D Command  
Code 44  
Naval Medical Center  
Bethesda, MD 20014

Director  
Behavioral Sciences Dep.  
Naval Medical Research Institute  
Bethesda, MD 20014

Dr. George Moeller  
Human Factors Engineering Branch  
Submarine Medical Research Lab.  
Naval Submarine Base  
Groton, CT 06340

Chief, Aerospace Psychology Division  
Naval Aerospace Medical Institute  
Pensacola, FL 32512

Navy Personnel Research and  
Development Center  
Management Support Dept.  
Code 210  
San Diego, CA 92152

Dr. Fred Muckler  
NPRDC  
Manned Systems Design, Code 311  
San Diego, CA 92152

Dr. Mel Moy  
Navy Personnel Research and  
Development Center  
Code 305  
San Diego, CA 92152

Human Factors Dept.  
Code N215  
Naval Training Equipment Center  
Orlando, FL 32813

Mr. J. Barber  
Headquarters, Dep. of the  
Army, DAPE-PBR  
Washington, D.C. 20546

Dr. Joseph Zeidner  
Acting Technical Dir.  
U.S. Army Research Institute  
5001 Eisenhower Ave.  
Alexandria, VA 22333

Dr. Edgar M. Johnson  
Organization and Systems  
U.S. Army Research Lab.  
5001 Eisenhower Avenue  
Alexandria, VA 22333

Technical Director  
U.S. Army Human Eng. Labs  
Aberdeen Proving Ground  
Aberdeen, MD 21005

ARI Field Unit - Leavenworth  
Post Office Box 3122  
Fort Leavenworth, Kansas 66027

U.S. Army Aeromedical Research  
Navy Personnel Research and  
Navy Personnel Research and Lab.  
ATTN: CPT Gerald P. Krueger  
Ft. Rucker, Alabama 36362

Naval Training Equip. Center  
ATTN: Technical Library  
Orlando, FL 32813

Dr. Alfred F. Smode  
Training Analysis and  
Evaluation Group  
Naval Training Equip. Center  
Code N-00T  
Orlando, FL 32813

U.S. Air Force Office of Scientific  
Research  
Bolling Air Force Base  
Washington, D.C. 20332

Dr. Donald A. Topmiller  
Chief, Systems Eng. Branch  
Human Eng. Division  
USAF AMRL/HES  
Wright-Patterson AFB, OH 45433

Lt. Col. Joseph A. Birt  
Human Eng. Division  
Aerospace Medical Research Lab.  
Wright Patterson AFB, OH 45433

Air University Lib.  
Maxwell Air Force Base, AL 36112

Dr. Robert Williges  
Human Factors Lab.  
Virginia Polytechnic Ins.  
130 Whittemore Hall  
Blacksburg, VA 24061

Dr. Meredith Crawford  
5606 Montgomery St.  
Chevy Chase, MD 20015

Dr. Terence R. Mitchell  
Univ. of Washington  
Seattle, WA 98195

Dr. Jesse Orlansky  
Ins. for Defense Analyses  
400 Army-Navy Dr.  
Arlington, VA 22202

Dr. William A. McClelland  
Human Resources Res. Office  
300 N. Washington St.  
Alexandria, VA 22314

Dr. Arthur I. Siegel  
Applied Psychological Services  
Life Sciences Directorate,  
NLInc.  
404 East Lancaster St.  
Wayne, PA 19087

Dr. Robert R. Mackie  
Human Factors Res., Inc.  
Santa Barbara Research Park  
6780 Cortona Dr.  
Goleta, CA 93017

Dr. Gershon Weltman  
Perceptronics, Inc.,  
6271 Variel Ave.  
Woodland Hills, CA 91364

Dr. Paul Slovic  
Decision Research  
1201 Oak Street  
Eugene, OR 97401

Dr. Alphonse Chapanis  
The Johns Hopkins Uni.  
Dep. of Psychology  
Charles & 34 Streets  
Baltimore, MD 21218

Dr. Clinton Kelly  
Decisions & Designs, Inc.  
8400 Westpark Dr., St. 600  
McLean, VA 22101

Dr. Melvin R. Novick  
Univ. of Iowa  
Lindquist Center for Meas.  
Iowa City, IA 52242

Journal Supplement Abstract  
Service  
American Psychological Asso.  
1200 17th St. N.W.  
Washington, D.C. 20036 (3  
cys)

Dr. Ron Howard  
Stanford Univ.  
Standford, CA 94305

AD-A060 786

OKLAHOMA UNIV NORMAN DECISION PROCESSES LAB  
HYPOTHESIS GENERATION AND PLAUSIBILITY ASSESSMENT. (U)  
OCT 78 C F GETTYS, S D FISHER, T MEHLE

F/G 5/10

UNCLASSIFIED

N00014-77-C-0615  
NL

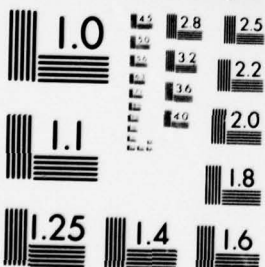
2 OF 2  
AD  
A060708



END  
DATE  
FILMED  
01-79  
DDC

OF 2

60786



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

Dr. Ward Edwards  
Dir., Social Science Res.  
Univ. of Southern Cal.  
Los Angeles, CA 90007

Dr. Miley Merkhofer  
Stanford Research Ins.  
Decision Analysis Group  
Menlo Park, CA 94025

Patricia A. Knoop  
AFHRL/ASM  
Wright-Patterson AFB OH 4

Director, Human Factors &  
Defence & Civil Ins. of  
Environmental Medicine  
Post Office Box 2000  
Downsville, Toronto, Onta  
Canada

Dr. David Zaidel  
Ins. for Research in Publ  
Univ. of Indiana  
Bloomington, IN 47401

Professor Judea Pearl  
Univ. of Cal.-Los Angeles  
Eng. Systems Dept.  
405 Hilgard Avenue  
Los Angeles, CA 90024

Edwards  
al Science Res. Ins.  
Southern Cal.  
es, CA 90007

Merkhofer  
Research Ins.  
Analysis Group  
c, CA 94025

A. Knoop  
ttersen AFB OH 45433

Human Factors Wing  
Civil Ins. of  
ntal Medicine  
ce Box 2000  
e, Toronto, Ontario

Zaidel  
Research in Public Safety  
Indiana  
on, IN 47401

Judea Pearl  
Cal.-Los Angeles  
ems Dept.  
rd Avenue  
es, CA 90024

Dr. Gary McClelland  
Univ. of Colorado  
Institute of Behavioral Sciences  
Boulder, CO 80309

Robert Carter  
Penn. State Uni.  
Psychology Dep.  
Moore Building  
Univ. Park, PA 16801

Prof. Howard Raiffa  
Grad. School of Bus. Adm.  
Harvard Univ.  
Soldiers Field Rd.  
Boston, MA 02163

Dr. A.D. Baddeley  
Director, Applied Psy. Unit  
Medical Research Council  
15 Chaucer Rd.  
Cambridge, CB2 2EF  
England

Prof. Dr. Carl Graf Hoyer  
Ins. for Psychology  
Technical University  
8000 Munich  
Arcisstr 21  
FEDERAL REPUBLIC OF GERMANY